Evaluation de modèles de classification appliqués à la détection d'opinions

Olena Zubaryeva, Jacques Savoy

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)
Olena.Zubaryeva@unine.ch, Jacques.Savoy@unine.ch

RÉSUMÉ. *Cet article présente et évalue différentes stratégies de classification automatique d'opinions. Ces dernières sont exprimées dans des phrases que le système doit classifier comme renfermant ou non une opinion. Dans ce but, nous avons retenu une classification basée sur le modèle* Naïve Bayes *et une autre basée sur des séparateurs à vaste marge (SVM). Comme alternative, nous suggérons un modèle basé sur le vocabulaire spécifique et le calcul d'un score normalisé (score Z). Au moyen de la collection test NCTIR, nos expériences démontrent que notre modèle apporte significativement la meilleure performance et que la représentation par des vocables s'avère préférable aux lemmes.*

ABSTRACT. *This paper describes the problem of classifying opinions expressed into sentences. The system must categorize them as opinionated or factual. To achieve this objective, we have used a* Naïve Bayes *approach and Support Vector Machines. As a new categorization model, we suggest using a normalized score (Z score) based on a specific vocabulary. Using a NTCIR test collection, our evaluations demonstrate that the suggested model based on the Z score performs significantly better than the others and that a representation based on words tends to show better performance level than surrogates based on lemmas.*

MOTS-CLÉS : *détection d'opinions; classification d'opinions ; évaluation de classifieurs.*

KEYWORDS: *Opinion detection, Opinion classification, Classifiers evaluation.*

1. Introduction

Avec le Web 2.0, les internautes ne se limitent plus à être des consommateurs d'information mais ils participent et collaborent à l'enrichissement des sites à l'image de l'encyclopédie Wikipédia. Mais cette fonction touche beaucoup d'autres secteurs via les blogs et les réseaux sociaux (e.g., FaceBook, Twitter). Sur ces divers supports, les usagers déposent leurs remarques, commentent les évènements ou les billets rédigés par d'autres. Ce rôle de producteur d'information permet à toutes les couches sociales d'exprimer et de faire partager ses sentiments, ses émotions ou ses opinions concernant toutes les activités humaines.

Si l'on consulte la blogosphère, la présence d'information de nature subjective apparaît de manière très visible, ce qui a conduit plusieurs moteurs commerciaux, à proposer une entrée distincte pour la recherche d'information (Boughanem & Savoy, 2008) dans les blogs. Dans un tel amalgame de commentaires et de jugements personnels, il serait intéressant de pouvoir extraire les éléments d'informations correspondant à une nature particulière comme, par exemple, les sentiments (bonheur, tristesse, colère, surprise, déception). Avec un tel système, nous pourrions connaître, par exemple, l'état d'esprit d'une population après un évènement particulier (e.g., attentats à Londres) (Mishne *et al*., 2006). Dans notre étude, nous nous sommes intéressés à distinguer entre des éléments objectifs (e.g., description d'un produit) et les phrases comprenant des opinions (positive, négative ou mixte).

La possibilité de distinguer, de manière assez fiable, entre les opinions et les données factuelles nous ouvre la voie à diverses applications intéressantes. Par exemple, un système de dépistage aurait la possibilité de proposer deux listes de réponse à la requête "batterie de l'iPhone". D'une part, on regroupe les éléments objectifs (e.g., poids, durée de vie, entretien, etc.) liés à la demande de l'internaute et, d'autre part, les opinions et jugements des internautes sur ce produit. En effet, les utilisateurs ou futurs acheteurs sont souvent plus sensibles aux expériences et avis d'autres personnes qu'à des informations purement factuelles. Dans d'autres domaines du traitement de la langue naturelle (Konchady, 2006), cette distinction peut trouver des perspectives d'application comme dans les systèmes de questions-réponses, la catégorisation automatique (Sebastiani, 2002), la génération de résumé ou le filtrage d'information (Minel, 2002).

La suite de cette communication se divise de la manière suivante. La section 2 expose un survol des travaux récents dans le domaine de la détection d'opinions tandis que les grandes lignes de la collection utilisée seront décrites dans la troisième section. La quatrième section discute des représentations choisies ainsi que des modèles de catégorisation sélectionnés. Notre nouveau modèle sera également discuté dans cette section. L'évaluation des divers modèles et représentations sera présentée et analysée dans la cinquième section.

2. Détection et classification d'opinions

Les billets déposés dans les blogs renferment souvent une opinion, des sentiments, des émotions ou, au contraire, peuvent correspondent simplement à une description. Cette discrimination soulève de nombreux défis en particulier lorsque l'on considère que la différence peut tenir sur un seul mot comme dans les phrases "The iPhone price is 600$" ou "The iPhone price is high". De plus, la distinction entre une opinion positive ou mixte peut parfois être sujette à interprétation et donc varier d'un individu à l'autre. Nous avons donc considéré plus simple de distinguer uniquement entre l'expression d'une opinion, sans considérer sa polarité pouvant être positive, négative ou mixte.

Malgré cette limitation, le problème de détection des opinions demeure difficile pour diverses raisons (Boiy & Moens, 2009). D'abord, nous devons reconnaître que toute langue naturelle autorise un très large éventail de possibilités lexicales, voire syntaxiques, pour exprimer la même idée (Furnas *et al*., 1987). De plus, l'écriture dans la blogosphère s'accompagne de contraintes moins fortes sur la forme comme sur la structure syntaxique. A la limite, on peut aboutir à un langage SMS (e.g., "iPhone KS"). La nature multilingue de la Toile et un nombre restreint d'exemples constituent des difficultés supplémentaires. Enfin, lorsque la propagande, la désinformation ou la manipulation interviennent le bruit devient trop important pour que ces opinions reflètent réellement des sentiments personnels.

Afin de déterminer si une phrase contient ou non une opinion, les systèmes de classification proposés se distinguent entre les approches s'appuyant d'une part sur le lexique et des connaissances symboliques (Esuli *et al*., 2006) ou, d'autre part, sur un apprentissage par machine (Abassi *et al*., 2008). Le résultat de plusieurs campagnes d'évaluation comme TREC (Macdonald *et al*., 2008) ou NTCIR (Seki *et al*., 2007), (Seki *et al*., 2008) démontrent l'intérêt pour ces deux types d'approche.

Dans le premier cas de figure, le système recherche des mots ou modèles caractéristiques (voire des parties du discours) afin de permettre une classification. Ainsi, Levin (1993) propose de définir différentes catégories verbales (déclaration, conjecture, 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). Dans ce type d'approche, on considère également les négations inversant la polarité de l'opinion ainsi que les cooccurrences fréquentes. Par exemple, on peut définir l'adjectif "beautiful" comme annonciateur d'une polarité positive. Si l'on retrouve également l'adverbe "very", on renforce la première direction tandis que la présence du mot "not" inversera cette direction (e.g., "not very beautiful", "not bad"). Turney (2002) constitue un exemple typique de cette école en proposant de considérer la fréquence d'apparition de termes et leurs cooccurrences et ceci conjointement avec un lexique contenant les mots annonçant des sentiments donnés.

La détection de phrases possédant ou non une opinion peut également être vue comme un problème d'apprentissage supervisé (Sebastiani, 2002). Dans ce cas, sur la base d'un ensemble d'exemples et de contre-exemples, on peut entraîner un classifieur afin qu'il produise le résultat escompté (Boiy & Moens, 2009). Dans cette voie de recherche, deux problèmes doivent être abordés (Abassi *et al*., 2008), à savoir d'une part comment représenter les documents ou phrases à analyser et, d'autre part, quel modèle de classification automatique utiliser. Durant les dernières campagnes internationales dans ce domaine, différentes équipes ont proposé de recourir au modèle *Naïve Bayes* (Mitchell, 1997), à des machines à vecteurs de support (Cristianini & Shawe-Taylor, 2006) ou à des modèles de langue (Pang *et al*., 2004). Au niveau des représentations, le recours à un "sac de mots" constitue le

socle sur lequel différentes variations sont possibles, comme la prise en compte de l'ordre des mots, la ponctuation ou la longueur des phrases. Comme exemple relativement complexe dans cet axe, on peut citer le système *OpinionFinder* (Wilson *et al*., 2005) permettant la détection d'opinion mais également une analyse de la nature subjective d'un écrit (e.g., sentiment, spéculation, rêve). La représentation se fonde sur les mots, mais également leur partie du discours (*parts of speech* ou POS). Un premier classifieur basé sur le modèle *Naïve Bayes* distingue les phrases objectives et subjectives. Cette distinction est également confirmée par un système à base de règles permettant de déceler l'annonce d'actes du discours ("said", "according to") ou l'expression de sentiments ("is happy", "fears").

Il est évident que la mise au point de tel système requiert un ensemble d'entraînement conséquent pour les approches basées sur un apprentissage par machine. Le volume insuffisant des exemples fût l'un des écueils majeurs rencontrés par toutes les équipes lors de la dernière campagne NTCIR-7. Par contre, les systèmes s'appuyant sur un lexique et des connaissances symboliques requièrent un travail manuel non négligeable dans la mise au point d'outils linguistiques adaptés (e.g., règles de production, thésaurus spécialisé). Finalement, ces deux approches peuvent se rejoindre dans la mise au point de systèmes hybrides ou dans le recours à l'apprentissage automatique afin de générer les connaissances symboliques requises.

3. Les campagnes d'évaluation NTCIR et mesures de performance

3.1. Descriptions des collections de test

Dans le but de promouvoir le développement d'outils automatiques pour le traitement des langues de l'Extrême-Orient et de l'anglais, les organisateurs des campagnes d'évaluation NTCIR ont dirigé durant trois ans une piste ayant pour objet la détection des opinions dans des phrases. Les campagnes 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. Une collection similaire écrite en langue japonaise ou chinoise traditionnelle est également disponible.

Les documents pertinents étant connus, les organisateurs ont 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. Mais l'expression d'une opinion n'est pas toujours 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. La phrase est jugée comme contenant une opinion si une majorité de jugements corrobore cette assertion (évaluation permissive ou tolérante).

Si l'on regroupe les deux corpus NTCIR-6 et NTCIR-7, nous disposons de 10 145 phrases dont 2 495 (ou 24,6 %) contiennent une opinion et 7 650 (ou 75,4 %) ne renferment pas d'opinion. La figure 1 reprend quelques exemples.

Sans opinion
<TEXT> Five years ago, there were no Internet-related information businesses.
Opinion négative
<TEXT> Since the United States is Korea's most important trade partner, the Korean economy was also affected immediately.
Opinion positive
<TEXT> ``I believe that we have found the appropriate balance,'' he said.
Opinion mixte
<TEXT> However, it is important that we not place excessively high expectations on the summit.

***Figure 1** : Quelques exemples de phrases du corpus d'évaluation*

3.2. Mesures d'évaluation

Durant ces diverses campagnes d'évaluation, la précision, le rappel et la mesure F ont été utilisés comme mesures de performance. La précision (formule 1) indique la probabilité qu'une phrase classifiée comme ayant une opinion possède réellement une opinion. Le rappel (formule 2) dénote la probabilité qu'une phrase ayant une opinion soit classifiée correctement. Ces deux mesures 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 correspondant à une moyenne harmonique et sera utilisée dans cette perspective. Dans le cas présent où $\beta = 1$, on accorde autant d'importance à la précision qu'au rappel.

$$\text{Précision } \pi = \frac{\#\,vrai\ \ positif}{\#\,vrai\ \ positif + \#\,faux\ \ positif} \tag{1}$$

$$\text{Rappel } \rho = \frac{\#\,vrai\ \ positif}{\#\,vrai\ \ positif + \#\,faux\ négatif} \tag{2}$$

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

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 sera trop optimiste et plus ou moins biaisée selon les modèles d'apprentissage (évaluation "rétrospective"). 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. Sur la base des k mesures de performance obtenues, nous pouvons calculer une moyenne. Notons que dans toutes nos évaluations, nous appliquerons exactement la même

subdivision en *k* blocs, les mêmes phrases apparaissant toujours dans les mêmes blocs. Sur cette base et pour définir s'il existe une différence statistiquement significative entre deux moyennes, nous pouvons appliquer le test bilatéral *t* pairé au seuil de signification $\alpha = 5\ \%$ (Grimm, 1993).

4. Modèle de catégorisation automatique

Tout modèle d'apprentissage doit disposer d'une représentation adéquate pour les documents ou phrases à classifier. Plusieurs questions doivent trouver une réponse appropriée comme l'élimination des mots-outils, l'application d'un enracineur ou la réduction des mots à des séquences de *n*-grammes (McNamee, 2009) ainsi que la pondération des éléments retenus. Ces questions seront abordées dans la section 4.1. tandis que les trois autres sections présenteront les trois modèles analysés.

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

Comme élément pertinent afin de représenter une phrase, 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. L'élimination des suffixes dérivationnels (Porter, 1980) peut être envisagée mais nous avons renoncé à cette solution dans cette étude. 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 léger, ou plus agressif ou le recours à un traitement morphologique, apporte des qualités de réponse statistiquement similaires (Fautsch & Savoy, 2009).

Comme l'usage du terme "mot" devient ambigu, nous avons réservé la désignation "vocable" pour indiquer les formes distinctes, tandis que le terme "mot" signalera une forme apparaissant dans un texte. Ainsi, dans la phrase "the dogs saw the brown dog", 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"), modification qui n'est pas toujours opportune (e.g., dans le nom "Steve Jobs"). Par contre, si le mot s'écrit uniquement avec des majuscules, celles-ci seront conservées (e.g., "US", "DOJ" ou "AIDS"). Comme d'autres langues naturelles, l'anglais connaît des variantes orthographiques comme "center", "centre" ou "defense", "defence" que notre système ne regroupera pas sous la même entrée.

Finalement, nous n'avons pas procédé à un traitement plus approfondi afin, par exemple, d'éliminer les suffixes dérivationnels (*e.g.*, “China” et “Chinese”). 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)).

Sur l'ensemble des 10 145 phrases de notre corpus, on compte 219 038 mots pour 14 025 vocables différents (ou 15 259 avant l'application d'un enracineur léger). 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 14 025 à 5 021 vocables, soit une réduction de 64,2 % (ou de 13 160 lemmes à 4 652 (-65,8 %)). L'élimination des vocables peu fréquents correspond d'abord à un souci de réduire sensiblement l'espace de représentation de nos classifieurs.

Afin d'avoir une idée des formes les plus fréquentes, le tableau 1 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 (colonne *tf*) ainsi que le nombre de phrases ayant au moins une occurrence du vocable (colonne *df*).

	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 1** : Vocables les plus fréquents dans les phrases renfermant une opinion (2 495 phrases) et sans opinion (7 650 phrases)*

4.2. Le classifieur 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 devra choisir entre deux hypothèses possibles soit h_0 = “sans opinion” et h_1 = “avec opinion”. La catégorie sélectionnée sera celle qui retournera

la valeur maximale de la formule 4. Dans cette dernière, t indique le nombre de vocables inclus dans la phrase courante et t_j les termes apparaissant dans la phrase.

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

Les probabilités sous-jacentes doivent être estimées. Pour les probabilités *a priori* 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 T_{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 T_{hi} (notée tf_{hi}) et la taille de l'ensemble correspondant.

$$\text{Prob}[t_j \mid h_i] = tf_{hi} \,/\, n_{hi} \tag{5}$$

Cette estimation (maximum de vraisemblance) 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 (Manning & Schütze, 2000). 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 $|V|$ la taille du vocabulaire (*e.g.*, 4 135 vocables si h_0 et 2 095 pour h_1).

4.3. Séparateurs à vaste marge (SVM)

La représentation d'une phrase peut se fonder sur la présence ou l'absence de termes avec une pondération reflétant 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 essentiellement au logarithme de l'inverse de la fréquence documentaire (notée *df*). Cette dernière valeur indique le nombre de phrases dans lesquelles ce terme apparaît, avec *n* désignant 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 $atf = 0{,}5 + 0{,}5 \cdot (tf / \max tf)$. 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), (Cristianini & Shawe-Taylor, 2006). 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. La prise en compte, dans ce modèle, de fonctions noyaux polynomiales ou sigmoïdales (transformant l'espace de représentation des vocables (pondération *tf idf*)) permet parfois d'accroître la performance au prix d'une augmentation sensible du temps de traitement.

4.4. Le score normalisé Z

Comme alternative, nous suggérons de pondérer les vocables en fonction de leur appartenance au vocabulaire spécifique (Muller, 1992) des phrases avec ou sans opinion. Afin de mesurer cette spécificité, on subdivise notre corpus en deux, soit entre les phrases sans opinion (ensemble noté h_0) et celles avec une opinion (ou h_1). Pour un terme ω, on compte le nombre d'occurrences dans l'ensemble h_0 (valeur notée tf_{h0}) et sa fréquence dans h_1 (notée tf_{h1}). Dans le corpus NTCIR, nous aurons tf_{h0}+tf_{h1} occurrences de ce vocable. La taille de l'ensemble h_0 s'élève à n_{h0} tandis que le volume du corpus sera n (= n_{h0} + n_{h1}).

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_{h1}. Prob[ω] désigne la probabilité de tirer le vocable ω lorsque nous tirons un mot au hasard dans le corpus. Cette probabilité s'estime par (tf_{h0}+tf_{h1}) / n. Si l'on répète n_{h1} fois ce tirage aléatoire, on peut estimer le nombre de vocables ω tiré par Prob[ω]·n_{h1}. Ce nombre correspond au nombre attendu que nous comparons avec le nombre observé, soit tf_{h1}. De manière précise, nous calculons un score Z pour chaque vocable ω selon l'équation 6 dans laquelle Prob[ω]·n_{h1} représente la moyenne et n_{h1}·Prob[ω]·(1-Prob[ω]) la variance de la binomiale.

$$score\ Z(\omega) = \frac{tf_{h1} - n_{h1} \cdot \text{Prob}[\omega]}{\sqrt{n_{h1} \cdot \text{Prob}[\omega] \cdot (1 - \text{Prob}[\omega])}} \quad (6)$$

Comme règle de décision, on peut considérer que les vocables présentant un score Z supérieur à un seuil δ donné correspondent à un suremploi et ceux ayant un score inférieur à -δ à un sous-emploi. Pour les phrases ayant une opinion, les vocables possédant les plus fortes valeurs Z sont “should” (score Z = 7,17), “we” (4,95), “must” (4,7), “because” (3,79) ou “right” (3,75). Inversement, dans les sous-emplois ou les vocables sur-employés dans les phrases sans opinion, on retrouve “year” (score Z = 5,96), “last” (4,03), “billion” (3,95), “police” (3,85) ou “first” (3,76). En recourant à cette méthode, nous avons mis en lumière les différences de vocabulaire dans les discours électoraux suisses ou français (Savoy, 2009).

---

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

Afin de déterminer si une phrase contient une opinion, on récupère le score Z de chaque vocable ou lemme 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 (h_1), sinon dans celle des phrases sans opinion (h_0).

5. Evaluation

5.1. Performance des différents classifieurs

Sur la base du corpus NTCIR, 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 2. Dans celui-ci, nous avons recouru à plusieurs représentations par modèles comprenant des vocables ou des lemmes, en utilisant (symbole « λ ») ou non un lissage des probabilités ou en éliminant les termes ayant une fréquence inférieure à 4 (« min : 4 »).

En comparant les meilleures performances obtenues par les trois classifieurs, notre modèle, basé sur le score Z, propose la meilleure stratégie et ceci en considérant les trois mesures retenues. En tenant compte uniquement de la mesure $F_{(1)}$, on constate que la différence de performance est nette entre le score Z (nº 16) et l'approche SVM (nº 10) (57,34 % vs. 45,33 %, soit une différence relative de 21 %). Avec le modèle *Naïve Bayes*, la plus forte valeur $F_{(1)}$ se situe à 30,49 % (nº 6), soit une différence relative de 46,83 % avec le score Z (nº 16).

Afin de confirmer cette conclusion, nous pouvons appliquer le test *t* pairé (bilatéral, niveau de signification $\alpha = 5$ %). Sur la base de la représentation à l'aide de vocables (avec lissage λ et min : 4), la performance selon la mesure $F_{(1)}$ du score Z (nº 13) s'avère significativement différente de celle du modèle SVM (nº 9) ou *Naïve Bayes* (nº 1, différence signalée par le symbole « † » dans le tableau 2). Sur la base du rappel, nous obtenons la même confirmation, tandis que pour la précision la différence de performance entre les modèles score Z et SVM n'est pas significative. Si l'on base nos représentations sur des lemmes (avec lissage λ, min: 4 soit nº 17 avec nº 11 et 5), les différences de performance entre le modèle fondé sur le score Z et les deux autres sont statistiquement significatives, que l'on utilise la précision, le rappel ou la mesure $F_{(1)}$.

Pour connaître la représentation proposant la meilleure performance, les valeurs du tableau 2 indique clairement que l'emploi de vocables s'avère plus approprié que les lemmes pour le classifieur basé sur le score Z. Cependant, pour le modèle *Naïve Bayes*, cette distinction n'apporte pas une variation sensible et systématique des performances. La situation s'avère assez similaire avec le modèle SVM. La représentation par vocables améliore la précision mais détériore quelque peu le rappel. Les différences de performance entre ces représentations restent faibles.

Afin de réduire sensiblement l'espace de représentation (de l'ordre de 60 %), nous avons suggéré d'ignorer les termes ayant une fréquence inférieure à quatre. De

plus, nous pouvons lisser les probabilités sous-jacentes selon l'approche de Lidstone (avec $\lambda = 0{,}3$). Ces techniques ne possèdent pas un effet systématique sur la performance au regard des différents classifieurs. Nous pouvons cependant dégager quelques tendances. Les variations de performance restent faibles entre les représentations avec tous les termes ou seulement ceux ayant une fréquence supérieure à trois. Réduire l'espace des représentations tend à favoriser la précision du modèle basé sur le score Z (e.g., nº 13 vs. nº 14) ou à augmenter le rappel des approches *Naïve Bayes* (e.g., nº 1 vs. nº 2). Dans ce dernier cas, signalons que de tenir compte de tous les termes et d'ignorer le lissage des probabilités sous-jacentes détériorent nettement les performances (e.g., nº 4 ou nº 8).

Nº	Modèle, paramètres	Précision	Rappel	$F_{(1)}$
1	Naïve Bayes, vocable, λ, min : 4	18,89 % †	67,45 % †	29,52 % †
2	Naïve Bayes, vocable, λ, min : 0	19,50 %	61,48 % *	29,60 %
3	Naïve Bayes, vocable, min : 4	19,05 % *	68,09 % *	29,76 % *
4	Naïve Bayes, vocable, min : 0	13,20 % *	37,08 % *	19,46 % *
5	Naïve Bayes, lemme, λ, min : 4	18,21 % †	63,07 % * †	28,26 % †
6	Naïve Bayes, lemme, λ, min : 0	20,15 %	63,03 % *	*30,49 %*
7	Naïve Bayes, lemme, min : 4	18,67 %	64,80 % *	28,99 %
8	Naïve Bayes, lemme, min : 0	14,64 % *	42,01 % *	21,69 % *
9	SVM, vocable, *tf idf*	33,63 %	65,76 % †	44,37 % †
10	SVM, vocable, *atf nidf*	34,99 % *	64,97 %	*45,33 %*
11	SVM, lemme, *tf idf*	32,42 % †	66,80 % †	43,33 % †
12	SVM, lemme, *atf nidf*	33,06 %	67,19 %	43,95 %
13	Score Z, vocable, λ, min : 4	44,23 %	82,72 %	56,30 %
14	Score Z, vocable, λ, min : 0	43,93 %	**84,49 %** *	56,50 %
15	Score Z, vocable, min : 4	**45,54 %** *	81,00 % *	57,01 % *
16	Score Z, vocable, min : 0	45,40 % *	82,97 %	**57,34 %** *
17	Score Z, lemme, λ, min : 4	39,29 %	81,71 %	52,87 %
18	Score Z, lemme, λ, min : 0	39,19 %	83,80 %	53,23 %
19	Score Z, lemme, min : 4	41,70 %	77,38 % *	53,92 %
20	Score Z, lemme, min : 0	41,46 %	79,91 % *	54,36 %

***Tableau 2** : Evaluation de diverses stratégies de catégorisation (validation croisée, 10 blocs; 2 495 avec opinion, 7 650 sans)*

Afin d'identifier les différences importantes entre les diverses représentations, chaque modèle débute par un classifieur de référence (soit nº 1 (Naïve Bayes), nº 9 (SVM) ou nº 13 (score Z)). Les différences de performance statistiquement significatives avec ce modèle sont indiquées par le symbole « * » dans le tableau 2.

Enfin, nous étions convaincu que la valeur du paramètre λ dans l'estimation des probabilités n'avait qu'un impact mineur dans la mesure de performance de notre modèle basé sur le score Z. Les valeurs de la mesure $F_{(1)}$ indiquées dans le tableau 3 confirme cette hypothèse. Ces dernières sont calculées sur la base du modèle n° 13 (voir tableau 2) avec, à la limite si λ = 0, les valeurs de performance obtenues par le modèle n° 15 ($F_{(1)}$: 57,01 %). Une amélioration substantielle de la performance n'est donc à rechercher dans un ajustement de ce paramètre sous-jacent.

λ = 0,001	λ = 0,1	λ = 0,3	λ = 0,5	λ = 0,75	λ = 1,0
56,95 %	56,70 %	56,30 %	56,05 %	55,77 %	55,39 %

***Tableau 3** : Évaluation $F_{(1)}$ (validation croisée) avec diverses valeurs pour le paramètre λ (modèle Score Z, vocable, min : 4)*

5.2. Variations du modèle SVM

Pour le modèle SVM, on constate que les variations des pondérations *tf idf* ou *atf nidf* ne modifient pas sensiblement la performance. De plus, le recours à des fonctions noyaux sigmoïdales transformant l'espace de représentation des vocables (pondération *atf nidf*) améliore quelque peu la performance ($F_{(1)}$: 46,26 % vs. 45,33 %, différence relative de 2,05 %). Cette augmentation se réalise par une amélioration de la précision (40,93 % vs. 34,99 %), accompagnée d'une baisse du rappel (54,53 % vs. 64,97 %). Par contre, le temps de traitement augmente de l'ordre de 2 000 % limitant l'intérêt pour des représentations plus complexes. Comme alternative, nous pouvons faire varier le paramètre C (coût) marquant la tolérance aux erreurs dans l'ensemble d'entrainement. Ainsi une faible valeur de C indique une plus grande tolérance aux erreurs dans l'ensemble d'apprentissage (marge souple) tandis qu'une valeur importante pénalisera fortement de telles erreurs.

Comme variations possibles avec notre modèle SVM de base (n° 9, tableau 2, $F_{(1)}$: 44,37 %), nous avons repris les valeurs suggérées par Joachims (2002) et les valeurs $F_{(1)}$ correspondantes sont indiquées dans le tableau 4. Avec le meilleur choix (C = 0,05), la meilleure performance passe à 50,19 % (soit un accroissement relatif de 13,1 %). Contrairement à notre score Z (tableau 3), le choix de la valeur du paramètre C possède un impact important, sans que l'on puisse, a priori, connaître une valeur optimale (e.g., Joachims (2002) signale une valeur C = 5 dans un système de classification de nouvelles du corpus *Reuters* ou de la collection *Ohsumed*[2]).

Défaut	C = 0,001	C = 0,05	C = 0,1	C = 1	C = 10	C = 1000
44,37 %	49,58 %	50,19 %	49,34 %	48,06 %	47,22 %	44,18 %

***Tableau 4** : Évaluation $F_{(1)}$ (validation croisée) avec diverses valeurs pour le paramètre C (modèle SVM, vocable, tf idf)*

[2] A l'adresse http://www.daviddlewis.com/resources/testcollections/reuters21578/ on retrouve le corpus *Reuters* tandis que la collection Oshumed est disponible à ftp://medir.oshsu.edu/pub/ohsumed

Finalement, les différences entre une évaluation rétrospective (même ensemble pour l'entraînement et l'évaluation) et la validation croisée s'avère élevées. Ainsi, avec une pondération *tf idf* et une représentation par des vocables, la mesure $F_{(1)}$ passe de 44,37 % à 71,77 %, soit une augmentation relative de 61,7 %. 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'en trouve nettement facilitée et la mesure de performance fortement biaisée.

5.3. Analyse de quelques phrases

Afin d'expliquer les différences de décision entre les classifieurs, nous avons analysé quelques représentations et erreurs de classification. L'approche *Naïve Bayes* possède, à nos yeux, l'inconvénient de ne pas discriminer fortement les vocables en fonction de leur présence plus ou moins importante dans l'ensemble des phrases avec opinion ou sans opinion. En fait, l'estimation indiquée dans la formule 5 tient compte uniquement de la fréquence d'occurrence dans la catégorie considérée. Ainsi, les trois vocables possédant les plus fortes probabilités sont identiques dans les deux catégories (voir tableau 1). Il s'agit de “said” (avec opinion : $16{,}83 \cdot 10^{-3}$, sans opinion : $8{,}11 \cdot 10^{-3}$), “not” (avec opinion : $13{,}25 \cdot 10^{-3}$, sans opinion : $6{,}79 \cdot 10^{-3}$) ou “he” (avec opinion : $9{,}11 \cdot 10^{-3}$, sans opinion : $5{,}80 \cdot 10^{-3}$).

Dans le tableau 5, nous avons repris quelques phrases provoquant des erreurs de classification pour l'un ou l'autre de nos modèles. Dans chaque cas, nous avons donné la catégorie ainsi que le texte de la phrase (<TEXT>). Après l'étiquette <NB>, nous avons regroupé la représentation selon le modèle *Naïve Bayes*. Dans ce cas, on a indiqué entre parenthèses les probabilités d'appartenir à la catégorie des phrases avec opinion et sans opinion puis, pour chaque terme, les deux mêmes probabilités (multipliée par 10^5). Pour le modèle SVM, nous avons repris la pondération *tf idf* calculée pour chaque vocable. Enfin, pour le score Z, nous avons signalé chaque mot suivi de sa valeur Z. Au début de cette représentation, nous avons donné, entre parenthèses, les valeurs *sumPos* et *sumNeg* qui s'utilisent pour décider si une phrase appartient à la catégorie des phrases avec ou sans opinion.

Dans le tableau 5, la première phrase s'avère difficile à classifier correctement car un terme important “psychiatry” possède une fréquence unitaire et est donc ignorée (présence du symbole “-”). La décision finale dépend exclusivement du vocable “half” poussant le modèle *Naïve Bayes* ou celui basé sur le score Z vers la catégorie “sans opinion”.

Dans la deuxième phrase, les modèles *Naïve Bayes* et score Z aboutissent à une décision incorrecte. Dans ce cas, les deux vocables “south” et “korean” font pencher la balance vers la classe “sans opinion” sans que les autres termes puissent contrer cet effet (e.g., les vocables “clearly” ou “say”). A nouveau, des termes pouvant être importants disposent d'une fréquence lexicale trop faible pour influencer la décision (e.g., “provocative” avec une fréquence lexicale de trois).

Opinion mixte	
<TEXT> Half of the job is psychiatry	
<NB> (0,179 / 0,821)	half (3,23 / 5,91) job (2,61 / 2,13) psychiatry (-)
<SVM>	half (5,05) job (5,93)
<score Z> (0,0 / -1,83)	half (-1,83) job (0,16) psychiatry (-)
Opinion négative	
<TEXT> The hawks say the sub's mission was provocative, as it clearly infringed on South Korean waters.	
<NB> (0,365 / 0,635)	hawk (-) say (10,76 / 6,96) sub (1,35 / 0,77) mission (2,92 / 2,03) provocative (-) clearly (4,49 / 1,19) infringed (-) south (23,32 / 33,11) korean (32,42 / 38,68) water (2,29 / 4,76)
<SVM>	hawk (7,43) say (4,63) sub (6.83) mission (5,89) clearly (6,0) south (3,31) korean (3,22) water (5,33)
<score Z> (5,18 / -6,87)	hawk (1,01) say (1,39) sub (0,56) mission (0,52) provocative (-) clearly (2,78) infringed (-) south(-2,95) korean (-2,05) water (-1,87)
Opinion négative	
<TEXT> You were often abused and humiliated.	
<NB> (0,397 / 0,603)	you (12,65 / 7,7) often (4,17 / 3,39) abused (-) humiliated (-)
<SVM>	you (4,64) often (5,42) abused (7,15)
<score Z> (1,76 / 0)	you (1,76) often (0,26) abused (-0,15) humiliated (-)

Tableau 5 *: Trois exemples de phrases dans divers modèles de catégorisation*

Finalement, la troisième phrase est correctement classifiée par le score Z mais pas par le modèle *Naïve Bayes*. Le terme "humiliated" sera ignoré car sa fréquence est unitaire (*hapax*). Le terme "abused" n'apparaissant que deux fois dans les documents avec une opinion sera ignoré dans la représentation de l'approche *Naïve Bayes*. Dans ce dernier cas, la décision doit être prise uniquement sur la base de deux vocables ("you", "often") favorisant la présence d'une phrase avec une opinion (dans le rapport de 2/1). Cependant, le rapport des probabilités a priori (1/4) est en faveur des phrases sans opinion ce qui, dans ce cas, fait pencher la balance vers une décision "phrase sans opinion".

6. Conclusion

Dans cet article, nous avons proposé plusieurs représentations possibles d'une phrase. Comme modèle de catégorisation, nous avons implémenté l'approche *Naïve Bayes* ou des séparateurs à vaste marge (SVM). Notre modèle, basé sur le score Z, repose sur une différenciation entre les distributions lexicales dans les phrases avec ou sans opinion.

Notre modèle basé sur le score Z permet d'apporter une qualité de réponse significativement supérieure aux autres approches. Dans ce cas et comme pour le modèle SVM, la représentation par des vocables s'avère meilleure que celle reposant sur des lemmes. Appliquer un lissage des probabilités lexicales permet d'accroître le rappel, mais à tendance à détériorer un peu la précision. Ignorer les termes ayant une fréquence lexicale trop faible (par exemple inférieure à quatre) permet de réduire sensiblement l'espace des représentations (de l'ordre de 60 %) sans présenter des variations importantes au niveau des performances.

Comme perspective, la combinaison de ces diverses approches pourrait 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. Toutefois, le nombre d'exemples disponibles (soit 2 495) demeure très limité pour permettre un apprentissage automatique de bonne qualité.

Remerciements

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

Bibliographie

Abassi A., Chen H. & Salem A. « Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums », *ACM-Transactions on Information Systems*, vol. 26, n° 3, 2008.

Baayen H.R. *Word Frequency Distributions*, Dordrecht, Kluwer Academic Publishers, 2001.

Bloom K., Stein S. & Argamon S. « Appraisal extraction for news opinion analysis at NTCIR-6 », *Proceedings of NTCIR-6*, Tokyo, 15-18 May 2007, NII, p. 279-289.

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

Boughanem M. & Savoy J. *Recherche d'information*, Paris, Hermès, 2008.

Cristianini N. & Shawe-Taylor, J.T. *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods,* Cambridge, Cambridge University Press, 2006.

Esuli A. & Sebastiani F. « SentiWordNet: A publicly available lexical resource for opinion mining », *Proceedings LREC-06*, Lisbon, 26-28 May 2006, p. 417-422.

Fautsch C. & Savoy J. « Algorithmic stemmers or morphological analysis: An evaluation », *Journal of the American Society for Information Sciences & Technology*, vol. 60, n° 8, 2009, p. 1616-1624.

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

Furnas D., Landauer T.K., Gomez L.M. & Dumais T. « The vocabulary problem in human-system communication », *Communications of the ACM*, vol. 30, n° 11, 1987, p. 964-971.

Grimm L.G. *Statistical Applications for the Behavioral Sciences*, John Wiley & Sons, 1993.

Harman D. « How effective is suffixing? », *Journal of the American Society for Information Science*, vol. 42, n° 1, 1991, p. 7-15.

Holte R. C. « Very simple classification rules perform well on most commonly used datasets », *Machine Learning*, vol. 11, nº 1, 1993, p. 63-90.

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

Konchady M. *Text Mining Application Programming*, Boston, Ch. River, 2006.

Levin B. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: The University Of Chicago Press, 1993.

Macdonald C., Ounis I. & Soboroff I. « Overview of the TREC-2007 blog track », *Proceedings TREC-2007*, Gaithersburg (MD), 18-21 November 2008, NIST Publication #500-274.

Manning C.D. & Schütze H. *Foundations of Statistical Natural Language Processing*, Cambridge (MA), The MIT Press, 2000.

McNamee P., Nicholas C. & Mayfield J. « Addressing morphological variation in alphabetic languages », *Proceedings ACM-SIGIR*, Boston, 19-23 July 2009, ACM Press, p. 75-82.

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

Mishne G. & de Rijke M. « MoodViews: Tools for blog mood analysis », *Proceedings of AAAI*, Boston, 16-20 July 2006, p. 153-154.

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

Muller C. *Principes et méthodes de statistique lexicale*, Paris, Honoré Champion, 1992.

Pang B. & Lee L. « A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts », *Proceedings of ACL-2004*, Barcelona, 21-26 July 2004, p. 271-278.

Porter M.F. « An algorithm for suffix stripping », *Program*, vol. 14, 1980, p. 130-137.

Savoy J. « Indexation et représentation comparative : Application au discours électoral », *Actes CORIA'09*, Toulon, 5-7 mai 2009, p. 185-200.

Sebastiani F. « Machine learning in automatic text categorization », *ACM Computing Survey*, vol. 14, nº 1, 2002, p. 1-27.

Seki Y., Evans D.K., Ku L-W., Chen H.-H. & Noriko K. « Overview of opinion analysis pilot task at NTCIR-6 », *Proceedings NTCIR-6*, Tokyo, 15-18 May 2007, NII, p. 265-278.

Seki Y., Evans D.K., Ku L-W., Sun L., Chen H.-H. & Noriko K. « Overview of multilingual opinion analysis task at NTCIR-7 », *Proceedings NTCIR-7*, Tokyo, 16-19 December 2008, NII, p. 185-203.

Toutanova K., Klein D., Manning C. & Singer Y. « Feature-rich part-of-speech tagging with a cyclid dependency network », *Proceedings of HLT-NAACL 2003*, Edmonton, 27 May – 2 June 2003, p. 252-259.

Turney P. « Thumbs up, thumbs down? Semantic orientation applied to unsupervised classification of reviews », *Proceedings of the ACL,* Philadelphia (PA), 6-12 July 2002, p. 417-424.

Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E. & Patwardhan, S. « OpinionFinder: A system for subjectivity analysis », *Proceedings HLT/EMNLP*, Vancouver (BC), 6-8 October 2005, p. 34-35.