Approches naïves à l'analyse d'opinion

Eric Crestan, Stéphane Gigandet et Romain Vinot

Yahoo! Inc., Paris,
{ecrestan,sg,romainv}@yahoo-inc.com

Résumé : L'analyse d'opinion semble être une tâche simple pouvant se rapporter au simple paradigme de classification. Cependant, cela s'avère plus compliqué notamment à cause de la classe d'opinion *neutre* qui peut être à la fois une combinaison de positif/négatif ou aucun des deux. Nous montrons dans cet article que des systèmes rudimentaires, donc l'un est basé sur des modèles SVM entrainés sur une sélection de traits saillants et l'autre, basé sur une sommation d'indices pondérés, obtiennent des scores honorables dépassant la moyenne des systèmes participants. De plus, il est à noter, que ces systèmes n'utilisent aucunes ressources externes.

Mots-clés : Analyse d'opinion, Support Vector Machine, Critère d'impureté de Gini

1 Introduction

L'analyse d'opinion est depuis longtemps une composante importante des systèmes de veille technologique et stratégique sur Internet. On conçoit aisément le besoin des grandes sociétés de surveiller leur image de marque dans la presse quotidien, mais également à travers les multiples blogs et forums disponibles sur Internet.

Le domaine de l'analyse d'opinion sur le Web a connu un essor plus important ces dernières années avec la possibilité d'obtenir des corpus déjà annotés à travers les sites de ventes en ligne. Les travaux de (Schein *et al.*, 2002; Kushal *et al.*, 2003) utilisent les revues de produits comme un tout afin d'entrainer des classifiers. D'autres systèmes considèrent des fenêtres de taille fixe autour d'entités nommées (Grefenstette *et al.*, 2001) ou encore, travaillent au niveau de la phrase (Wilson *et al.*, 2005). Cependant, la majorité de ces travaux ont été effectués sur la langue anglaise par manque de corpus annotés dans d'autres langues. La compagne DEFT'07 est donc la première à offrir un cadre d'évaluation pour l'analyse d'opinion sur une large diversité de type de documents en français.

Notre participation à cette campagne a été motivée par l'intérêt de connaître à quel point des systèmes simples peuvent performés sur une tâche supervisé d'analyse d'opinion. Les approches que nous proposons dans ces travaux sont basées, pour la première, sur les méthodes SVM couplées avec une sélection de traits fondée sur le saillance ; alors que notre second système se base sur la sommation des scores attribués aux couples trait/classe. Ces deux systèmes, qui reposent principalement sur l'heuristique et l'empirique, ne prétendent pas révolutionner le domaine, mais montrent qu'il est possible d'arriver à des résultats acceptables avec peu d'efforts.

Cet article est découpé en trois parties principales : Dans un premier temps, les deux approches seront présentées en détail. La seconde partie sera consacrée à l'évaluation de nos systèmes dans le cadre de la campagne DEFT'07. Enfin, nous terminerons cet article par une analyse détaillée des résultats en proposant des exemples concrets issus des jeux de test.

2 Description des approches

Les différents systèmes présentés dans cette section, sont basés sur des approches supervisées n'utilisant aucun lexique externe, si ce n'est une courte liste de mots-outil.

2.1 Approche SVM et sélection de traits

Les *Support Vector Machines* (SVM) sont devenus très populaires pour les systèmes de classification supervisée depuis leur application à ce domaine par (Joachims, 1997). Bien que cela soit une approche binaire, elle peut tout à fait être appliquée sur de grande dimension en utilisant un mode d'apprentissage *un contre tous* et en construisant donc un classifier par classe.

2.1.1 Application des SVM à l'analyse d'opinion

Dans le cadre de l'évaluation DEFT'07, la plupart des corpus sont basés sur une classification ternaire (*positive*, *négative* et *neutre*), ce qui nous oblige à entrainer plusieurs modèles. Seul le corpus de *débats parlementaires*, requière une classification binaire (*pour* / *contre*).

La logique en classification de document par SVM est d'avoir un classifier par classe. Toutefois, notre application est quelque que peu différente des schémas classiques car les classes représentent des tonalités et non des thèmes. De plus, il convient de s'interroger sur la sémantique des classes proposées. En effet, il semble difficile de trouver des termes décrivant la neutralité. Pour cette raison, seuls 2 classifiers seront créés, dont un servira à la détection des traits *positifs* (*positif* contre *négatif*+*neutre*), l'autre à la détection des traits *négatifs* (*négatif* contre *positif*+*neutre*). La classe *neutre* sera, quand à elle, affectée dans le cas ou ni le classifier *positif*, ni le classifier *négatif* n'auront un score dépasseront le seuil requit.

Dans le cadre de cette évaluation, nous avons mis en œuvre l'outil *mySVM* de (Stefan Rüping, 2000). La particularité de cet outil est qu'il n'accepte pas de grande dimension et nous avons donc dû nous contenter de 80 dimensions pour ces travaux.

2.1.2 Sélection des traits

Une des phases les plus importants lors de l'entrainement des SVM est la sélection des traits (ou *features*). La grande dimensionnalité qu'offrent les corpus, se comptant en milliers, ne permet pas une utilisation totale de ces dimensions. Il est donc indispensable de réduire la dimensionnalité en faisant une sélection des traits les plus porteurs d'information pour la tâche. De nombreuses approches ont été proposées par le passé, dont les plus populaires et performantes sont le test du χ^2 (Schütze *et al.*, 1995) et le gain d'information (Yang & Pedersen, 1997).

Pour notre part, nous proposons l'utilisation d'une variante du critère de divergence de Kullback–Leibler (Kullback & Leibler, 1951), ce critère est également appelé critère du gain d'information. En effet, une divergence trop importante entre la distribution d'un mot entre les classes, constitue un indice sur l'important de ce terme à discriminer une classe par rapport à une autre. Nous ferons référence par la suite à ce critère de divergence comme *score de saillance*, définit par :

$$S(t, C_i) = \left[P(C_i / t) - P(t)\right] \times \log\left(\frac{P(C_i / t)}{P(t)}\right) \qquad (1)$$

Le score de saillance peut donc être calculé pour chaque terme *t* appartenant à la classe C_i. Dans notre évaluation, le nombre de classe est de 2 (*positif* et *négatif*), le cas de la classe *neutre* n'étant pas considéré.

Finalement, seuls les 40 termes ayant les plus hautes saillances par classe seront retenus pour l'apprentissage des modèles, correspondant au final à un vecteur de 80 dimensions.

2.1.3 Prétraitement et Apprentissage

Avant même de pouvoir identifier les termes saillants inhérents à une classe, les documents doivent être segmentés. Cependant, n'utilisant pas de dictionnaire externe, une segmentation uniquement effectuée sur les unités lexicales peut éventuellement générer une perte d'information. Ceci est encore plus vrai dans le cadre de cette évaluation car les opinions *positives* ou *négatives* sont souvent exprimées par des négations (*ne … pas*) comportant donc plusieurs mots. Par exemple, le verbe *parvenir* à une toute autre signification si celui-ci est précédé de la particule *ne* ou pas. Pour cela, les scores de saillance sont

calculés pour les uni-grammes, ainsi que pour les bi-grammes. Les *termes* définis comme les plus saillants, sont donc en fait composés à la fois d'unis et de bi-grammes. Des exemples de termes saillants seront présentés dans la Section 4 de cet article.

Le corpus d'apprentissage est constitué de vecteurs de traits de 80 dimensions pour chaque document, dans lesquelles apparaissent des 1 ou des 0 selon la présence ou non du terme. Les modèles SVM peuvent ensuite être entrainés sur ces vecteurs.

Le décodage se fait très simplement en créant les vecteurs par la même méthode pour les documents de test. Ensuite, la classe obtenant le score supérieur à zéro le plus élevé est choisie. Si aucun des modèles SVM ne donnent un score positif, la classe *neutre* est alors choisie par défaut.

2.2 Approche par maximisation d'indices

L'approche exposée dans cette section est des plus triviales et ne prétend pas être reconnue comme *approche scientifique* à proprement parlé. Principalement basée sur l'observation, le système proposé ici ne reste qu'une ébauche étant donné le peu de temps que nous lui avons consacrée. Cependant, les résultats obtenus sont des plus encourageants et nous invite donc à pousser plus avant notre analyse afin d'établir scientifiquement ce qui a été créé empiriquement.

Cette approche consiste à sommer pour chaque terme d'un document, les scores représentant les gains d'information observés sur le corpus d'apprentissage. Cette approche avait déjà été employée par (Crestan, 2004) dans le cadre de l'évaluation en désambiguïsation sémantique SENSEVAL-3 et avait montrée des résultats comparables à d'autres approches comme les arbres de classification sémantique.

2.2.1 Présentation de l'approche

Le grand problème des approches comme les modèles SVM ou d'autres approches, est qu'il est indispensable de réduire la dimensionnalité de la tâche. Cela engendre généralement une perte d'information, qui est toutefois nécessaire par le fait que ces approches opèrent généralement une recherche d'optimal qui peut être très coûteux en temps de calcul. D'autres approches, comme les arbres de classification, divisent les populations à chaque nœud et créé de nouvelles distributions avec des densités plus faibles rendant la généralisation difficile.

Cette approche simple, part de ce dernier constat. Lors de la construction d'arbre de classification, une fonction est utilisée afin de calculer le gain « d'ordre » que va procurer la réponse à une question. Dans le cadre des arbres de classifications sémantique, les questions portes sur la présence ou l'absence de certain terme en contexte. Suivant la réponse obtenue pour chacun des documents de l'apprentissage, deux populations se découpent : l'une pour laquelle la réponse a été affirmative et l'autre pour laquelle la réponse a été négative. Deux types de fonctions de gain sont communément utilisés afin de calculer le gain qu'apporte une question : Le gain d'entropie et le gain en impureté de Gini. Dans la présente approche, nous utilisons le gain en impureté de Gini afin de calculer l'apport d'un terme pour décrire une classe donnée.

Le coefficient de Gini a été créé par le statisticien Corrado Gini (1912) afin de mesurer le degré d'inégalité de la distribution des revenus dans une population. L'application de ce critère d'impureté dans notre cas se traduit par la formule suivante :

$$G(Q) = 1 - \sum_{c \in C} P(c/Q)^2 \qquad (2)$$

où *P(c/Q)* est la probabilité de la classe *c* sachant la distribution *Q*.

De par cette formule, nous pouvons connaître l'impureté d'une distribution initiale. De même, il est possible de calculer le gain d'impureté porté par un terme en faisant la différence entre la moyenne pondérée du gain d'impureté de la population contenant le terme et de celle ne le contenant pas avec l'impureté initiale. Cela se traduit par la formule suivante :

$$GI(t) = G(Q) - \frac{|Q_t| G(Q_t) + |Q_{\neg t}| G(Q_{\neg t})}{|Q|} \qquad (3)$$

où $G(Q)$ est l'impureté de Gini pour la population initiale, $G(Q_t)$ est l'impureté de Gini pour la distribution de document contenant le terme t et $G(Q_{\neg t})$, celle qui ne le contient pas.

Chaque terme a un impact différent suivant la classe dans laquelle il est observé. Par exemple, le terme *insipide*, ne va pas avoir le même poids dans la classe *positive* que dans celle *négative*. Pour cela, il est nécessaire de d'attribuer un score à chaque terme par rapport à une classe. La combinaison suivante a été trouvée de façon empirique sur le corpus d'apprentissage :

$$S(c_i, t) = \frac{GI(t) \times Q_{\neg t} \times Q_{t,c_i}}{Q_t^2} \quad (4)$$

où $Q_{t,ci}$ est la distribution des documents contenant le terme t qui appartiennent à la classe c_i.

Cette approche nous a permis d'obtenir les meilleurs résultats sur le corpus d'apprentissage et c'est donc celle-ci qui a été retenu pour l'évaluation. Cependant, plusieurs filtrages ont mis en place afin de ne conserver que certain trait répondant à plusieurs critères de seuil, qui n'ont d'ailleurs pas été optimisés jusqu'à maintenant. Ce filtrage a été mis en place dans l'objectif de ne conserver que les termes ayant le plus d'important pour la prise d'opinion. Typiquement, seuls les termes ayant une fréquence intra-classe supérieure à 50% de leur masse totale, ont été conservés. De plus, un seuil supplémentaire a été appliqué afin d'éliminer les termes ayant un score $S(c_i,t)$ faible.

2.2.2 Prétraitement et Apprentissage

Les mêmes prétraitements ont été appliqués pour cette approche, que celle présentée dans la section précédente. Il n'y a pas « d'apprentissage » à proprement dit, juste un calcul de poids pour chacun des termes du corpus d'entrainement. En plus des unis et bi-grammes, quelques traits supplémentaires ont été ajoutés comme celui indiquant qu'un terme est présent dans les deux premières ou les deux dernières phrases du document, ainsi qu'un trait additionnel pour les termes ayant une fréquence supérieurs à 3 dans un même document.

Le décodage se fait très simplement en sommant sur chaque classe, le score des termes contenus dans les documents.

3 Evaluation

Les corpus d'évaluation sont au nombre de quatre et couvrent les mêmes domaines que les corpus d'apprentissage présentés précédemment. Le tableau suivant donne le nombre d'exemples par classe pour l'apprentissage et l'évaluation.

Table 1 - *Distribution des exemples d'apprentissage et d'évaluation selon les corpus*

	Corpus d'apprentissage				***Corpus d'évaluation***			
	Positif	**Neutre**	**Négatif**	***Total***	**Positif**	**Neutre**	**Négatif**	***Total***
Film/Livre	1150	615	309	*2074*	768	411	207	*1386*
Jeux	874	1166	497	*2537*	583	779	332	*1694*
Relecture	376	278	227	*881*	256	190	157	*603*
Débat	6899		10400	*17299*	4961		6572	*11533*

On remarquera que la distribution des fréquences par classe est assez bien respectée entre apprentissage et évaluation pour tous les corpus.

3.1 Scores

Lors de l'évaluation, chaque équipe avait la possibilité de présenter jusqu'à 3 systèmes différents. Pour notre part, nous n'en avons proposé que deux. Au lieu de calculer la précision et le rappel en ne tenant pas compte des classe, ceux-ci sont calculés classe par classe, puis ces scores sont moyennés pour donner une précision et un rappel moyen.

N'ayant donné qu'une seule suggestion de classe à chaque fois avec un indice de confiance de 1, nos scores pondérés et non-pondérés sont identiques. Pour cette raison, nous ne considéreront pas le premier cas dans notre analyse.

3.1.1 Exécution 1

Notre premier système utilisant des SVM avec une sélection des 80 traits *positifs/négatif* les plus saillants obtient des scores dans la moyenne des autres participants, comme le montre la colonne Δ de la Table 2.

Table 2 - Scores obtenus pour l'exécution 1

	Précision	**Rappel**	**F-Score**	**Δ**
Film/Livre	0.542	0.517	0.529	+0.029
Jeux	0.678	0.662	0.670	+0.009
Relecture	0.458	0.424	0.441	-0.030
Débat	0.692	0.616	0.652	+0.010

Le corpus de *Relecture* semble être celui présentant le plus de difficulté pour notre système, mais également pour les autres systèmes si l'on en juge par la F-Score moyen (47,1%). A l'opposé, le corpus de *Débat* obtient le meilleur score, ce qui n'est pas étonnant étant donné la bipolarité des notes à attribuer (*pour* ou *contre*).

3.1.2 Exécution 2

Notre second système basé sur une sommation des scores individuels d'appartenance à une classe de chaque terme qui compose les documents, a obtenu, de manière surprenante, de meilleurs résultats que notre première approche. Seul le score obtenu sur le corpus de revue de *Films* et *Livres* est très légèrement inférieur au précédent (Table 3).

Table 3 - Scores obtenus pour l'exécution 2

	Précision	**Rappel**	**F-Score**	**Δ**
Film/Livre	0.500	0.547	0.523	+0.022
Jeux	0.718	0.633	0.673	+0.013
Relecture	0.507	0.425	0.462	-0.008
Débat	0.715	0.691	0.703	+0.061

Le gain le plus significatif est à observer sur le corpus de *Débat* pour lequel notre système obtient un F-Score supérieur à 70% avec +5% par rapport à notre premier système et +6% par rapport à la moyenne des systèmes.

4 Analyse des résultats par système

Il est très difficile de comparer les systèmes en se basant uniquement sur les scores bruts, sachant que ceux-ci sont en fait une moyenne de scores par classe. De plus, il est intéressant de décomposer les scores selon les classes afin d'appréhender la difficulté de chaque tâche selon le corpus. Nous nous efforcerons dans cette section d'effectuer une analyse détaillée de chaque système en présentant des exemples issus directement des corpus de test et d'apprentissage.

4.1 Analyse du système à base de SVM : *Exécution 1*

L'aspect le plus important et déterminant pour cette approche est la sélection des traits à utiliser pour entrainer les modèles SVM. La restriction des 80 traits est une contrainte supplémentaire qui impose une grande rigueur pour cette phase de l'apprentissage. Il ne saurait être question de sélectionner ces termes au hasard ou même, de prendre les termes les plus fréquents car beaucoup d'entre eux n'ont qu'une fonction de support, ou pire encore, sont présents dans une portion descriptive du document n'ayant pas pour but de donner une opinion sur le sujet (comme par exemple les synopsies de films). Cependant, il faudrait pourvoir détecter automatiquement ce qu'il retourne de l'opinion et ce qui est purement narration, ce qui n'est pas une chose aisée à effectuer de manière automatique. De plus, certaine fois, le ton employé pour la narration peut être porteur d'information quand à l'opinion de la personne qui a rédigée le commentaire, surtout dans le cas du corpus de *Relecture* où l'auteur est obligé de faire une synthèse de lui-même.

La Table 4 contient une série de mots utilisés pour l'apprentissage des modèles SVM sur le corpus *Film/Livre*.

Table 4 - *Top des termes saillants pour le corpus de critique de Film/Livre*

Saillance négative		**Saillance positive**	
malheureusement	0.483	écriture	0.033
._malheureusement	0.368	chef-d'	0.025
spectateur	0.366	mots	0.024
acteurs	0.321	chef-d'_oeuvre	0.024
ennui	0.307	magnifique	0.021
se_contente	0.286	roman	0.020
contente	0.269	artiste	0.020
comédie	0.257	sens	0.019
bons	0.234	oeuvre	0.019
intérêt	0.230	plume	0.018
cette_production	0.228	porte	0.017
hélas	0.215	personnel	0.017
décevant	0.211	maître	0.017
ridicule	0.201	littéraire	0.017
réalisation	0.199	évidemment	0.016
._hélas	0.196	rêves	0.016
Insipide	0.182	auteur	0.016
…	…	…	…

Plusieurs observations peuvent être émisses à partir de cet exemple. Dans un premier temps, on remarquera un problème dû à l'utilisation des unis et bi-grammes dans le même modèle. Ainsi, le terme *malheureusement* semble être le mot le plus saillant pour représenté un avis *négatif*. Mais, ce même mot est également listé en seconde position lorsqu'il est précédé d'un point (ce qui établi ça présence en début de phrase). Cela remet quelque peu en cause condition d'indépendance des variables.

Autre remarque, cette liste paraît contenir des termes semblant tout à fait dénués de tonalité comme par exemple les mots *spectateur* et *acteurs*. La raison en est que leur distribution au sein de la classe *négative* est incohérente avec celle de l'ensemble du corpus. Apparaissant dans 100 documents *négatifs* sur les 309 que comporte cette classe (environ 32% des critiques *négatives*), le mot *spectateur* à une distribution « diverge » par rapport à l'observation faite sur l'ensemble du corpus car n'apparaissant que 324 dans l'ensemble du corpus sur 2974 documents (donc, dans moins de 15% du corpus). A noter qu'aucune lemmatisation n'a été effectué sur les données, donc pluriels et singuliers sont traités comme des mots différents. Bien que ce mot ne soit pas teinté d'opinion, il est néanmoins une sorte de marqueur indiquant une prise de position souvent *négative* comme le montre les exemples suivant :

...s'abîme dans des considérations besogneuses et installe le ***spectateur*** *dans une torpeur sourde et désagréable.*

Réussite totale sur ce point, mais la formule choisie, véritable carcan, emprisonne le ***spectateur****.*

...la complaisance de cette chronique désenchantée épuisent les résistances du ***spectateur****.*

Le critique, prend souvent le *spectateur* à partie en décrivant les effets *négatifs* que tel ou tel film à sur lui ou elle. Ces termes ne sont toutefois pas porteurs d'opinion à eux seuls et bien qu'il augment la probabilité d'être face à une opinion *négative*, ils doivent cependant être secondés pas d'autres indices afin de réellement discriminer une classe par rapport à une autre. C'est d'ailleurs ce que nous pouvons observer à travers le poids qu'attribut l'apprentissage SVM à ces termes. Par exemple, le terme *spectateur* à un poids de 0.279, ce qui est positive, mais bien en dessous du *prior* qui est de -0,833. Donc, un document ne contenant que ce terme là, ne sera pas considéré comme *négatif* par le modèle.

Sur les 1386 documents que contient le corpus de test *Film/Livre*, 46 documents ne contenaient aucun des 80 termes sélectionnés lors de la construction des modèles. Le nombre de traits moyen non-nul pour ce même corpus est de 4,2 termes. Ce qui dénoté qu'en moyenne 4 termes sur les 80 sélectionnés ont été utilisés lors du décodage pour les documents de test. Ce chiffre passe à 5,3 pour le corpus de *Jeux* (et 28 documents vides sur 1694), 5,6 pour les *Relectures* (et 40 documents vides sur 603) et 0,97 pour le corpus de Débats (et 6068 documents vides sur 11533). On remarquera que plus de la moitié des documents du test sur les textes de *Débat*, ne contiennent aucun des termes sélectionnés. Cela veux implique donc que la moitié des jugements sont données grâce au *prior*, qui est de 1,0 pour le modèle *négative*, donc une opinion *négative* par défaut. Ce qui est bien en accord avec la distribution des éléments *positif/négatif* dans le corpus où les exemples *négatifs* sont plus nombreux que ceux *positifs*.

Enfin, il convient d'analyser la capacité du système à détecter plus une classe par rapport à une autre. La **Fig. 1** présente les F-Scores pour chaque corpus et chaque classe. Le corpus *Débat* étant bipolaire, il ne comprend pas de score pour la classe *neutre*.

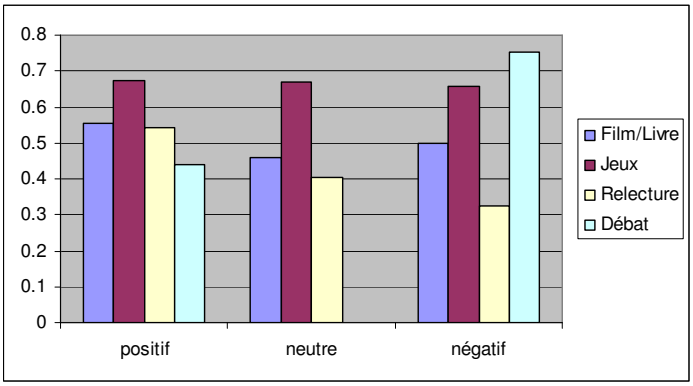

Fig. 1 - *F-Scores par classe et par corpus pour l'exécution 1*

Alors que le corpus de *Jeux* concède un score quasiment identique sur les trois classes, ce n'est pas le cas pour les autres corpus. Ainsi, on observe les plus grandes variations pour les corpus *Relecture* et *Débat* avec une préférence inversée quand à la classe la mieux identifiée. Cependant, il est difficile d'effectuer une analyse fine en utilisant le F-Score qui est une combinaison de précision et rappel. La Table 5 donne donc les scores de précision et de rappel pour chaque classe et pour chaque corpus de l'évaluation.

Table 5 - *Précision et Rappel par classe et par corpus pour l'éxécution 1*

	Précision			**Rappel**		
	positif	**neutre**	**négatif**	**positif**	**neutre**	**négatif**
Film/Livre	0.733	0.356	0.536	0.444	0.642	0.464
Jeux	0.745	0.630	0.658	0.617	0.710	0.660
Relecture	0.621	0.327	0.427	0.480	0.532	0.261
Débat	0.744	/	0.639	0.313	/	0.919

On observera en premier lieu que la précision de la classe *positive* est toujours la plus élevée que celle de la classe *négative*. Cela semble donc indiquer que ce système à tendance à n'affecter la classe *positive*

que lorsque la certitude est élevée. Par contre, la classe *neutre* semble être la classe par défaut pour tous les corpus à trois états et la classe *négative* pour le corpus *Débat*.

Cela semble traduire en fait l'incapacité du système à généraliser afin d'avoir assez de matière pour effectuer la classification. Le nombre réduit de traits utilisés pénalise grandement les classes d'apprentissage des SVM, car le jugement doit généralement se faire sur trop peu de termes.

4.2 Analyse du système à sommation de score : *Exécution 2*

Cette méthode, très empirique, à l'avantage de fonder sa décision sur un nombre beaucoup plus important de termes et donc d'indices. Par exemple, *1586* indices ont été extraits du corpus d'apprentissage *Film/Livre*. Ce nombre d'indice est très variable selon les corpus, il est de *5531* pour le corpus de *Jeux*, *1942* pour le corpus de *Relectures* et seulement *531* pour le corpus de *Débats*. Cette variation est due à plusieurs facteurs dont l'un des principaux est la mise en dur des seuils de filtrage dans le système.

La Table 6 présente des exemples de trait avec leur score respectif sur le corpus *Film/Livre*. Les traits précédés de *#LAST* et *#FIRST* correspondent à une présence de ces termes dans les deux dernières ou premières phrases du document. *#HF* indique, quant-à lui, une fréquence d'apparition supérieure ou égale à trois.

Table 6 - *Exemples de traits les plus saillants pour le corpus Film/Livre*

Négatif		**Neutre**		**Positif**	
Trait	**Score**	**Trait**	**Score**	**Trait**	**Score**
#LAST:ennuyeux	0.8150	ses_élèves	0.5193	oubli	0.1860
mollement	0.7697	agréable_à	0.5193	autre_côté	0.1764
remake#HF	0.7140	narre	0.4852	poignante	0.1692
lourdingue	0.7140	petite-fille	0.4852	#FIRST:chef-d'	0.1651
fade_.	0.7140	assez_bien	0.4852	plus_beau	0.1651
gâché_par	0.7140	est_dommage	0.4852	avancer	0.1588
consternant	0.7140	jeune_public	0.4852	#FIRST:magnifique	0.1579
décevant_.	0.6810	prof_de	0.4852	#LAST:photographie	0.1579
molle	0.6439	#LAST:jolis	0.4436	après-guerre	0.1579
vite_dans	0.6439	sérieuses	0.4436	#LAST:vérité	0.1553
scènes_sont	0.6439	face_.	0.4436	dépouillé	0.1533
insignifiante	0.6439	#LAST:en_attendant	0.4436	sans_oublier	0.1519
scène_sans	0.6439	#LAST:attendant	0.4436	plus_vite	0.1519
médiocrité_.	0.6439	#LAST:en_face	0.4436	pleines	0.1482
navet	0.6215	nos_héros	0.4436	narratrice	0.1482
ennuient	0.5536	#LAST:au_charme	0.4436	névroses	0.1482
…	…	…	…	…	…

Comme cela était le cas dans l'approche par sélection de traits pour les modèles SVM, tous les traits présentés ici ne semble pas tous pertinents par rapport à la classe. De plus, la description de la classe *neutre* est très abstraite dans le sens où il est difficile de définir la notion de neutralité pour un trait.

Le comportement du deuxième système est assez différent du premier pour certain corpus si l'on considère les *F-Scores* pour chaque classe. Ainsi, comme nous le montre la **Fig. 2**, la classe *neutre* obtient de moins bons scores, saufs pour le corpus de *Relectures*.

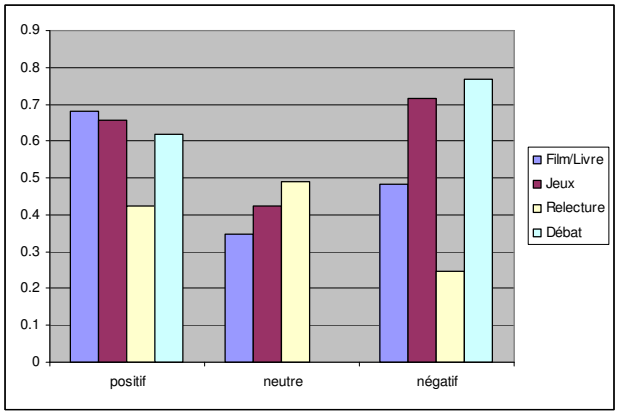

Fig. 2 - *F-Scores par classe et par corpus pour l'exécution 2*

Cela traduit directement la difficulté de la tâche de modélisation de cette classe. Par contre, la classe *positive* obtient de meilleurs scores (sauf pour le corpus de *Relectures*). Comme cela a été fait pour le système précédent, il faut analyser la précision et le rappel au niveau de chaque classe afin d'avoir une image claire de l'impact du système. La Table 7 nous montre une bonne précision du système en classe *positive*, sauf pour le corpus de *Jeux* pour lequel les classes *neutre* et *négative* dégagent une forte progression au détriment de la classe *positive*.

Table 7 - *Précision et Rappel par classe et par corpus pour l'éxécution 2*

	Précision			**Rappel**		
	positif	**neutre**	**négatif**	**positif**	**neutre**	**négatif**
Film/Livre	**0.708**	0.419	0.374	0.659	0.297	**0.686**
Jeux	0.490	0.826	**0.838**	**0.990**	0.286	0.623
Relecture	**0.696**	0.351	0.473	0.305	**0.805**	0.166
Débat	**0.721**	/	0.708	0.541	/	**0.842**

En regardant le nombre de traits utilisés afin de décrire chaque classe, on s'aperçoit qu'il y a une corrélation forte avec la précision sur cette classe. Les classes ayant le plus grand nombre de traits sont celles qui obtiennent les moins bons scores en précision. Par exemple, 54,3% des traits sélectionnés pour le corpus de *Jeux* sont dédiés à la classe *positive*. Un déséquilibre trop important dans le nombre de traits pour une classe plébiscite grandement celle-ci au détriment des autres, d'où une précision plus faible (peut être perçu comme la classe pas défaut).

Bien que ce système donne de meilleurs résultats que le précédent, il reste cependant très imparfait de par ses aspects empiriques. Il demande notamment un ajustement des paramètres de filtrage au niveau de chaque corpus. Pour cela, une technique de *N-Folds* devrait permettre une plus grande souplesse d'adaptation.

5 Conclusion

Nous avons montré à travers cette évaluation que des « systèmes simples », peuvent obtenir des résultats honorables, dans la moyenne des autres participants. Notre système à base de SVM et de sélection de traits basée sur la saillance des termes, permet d'atteindre une bonne précision sur la classe *positive* avec des scores supérieurs à 60%. Cependant, cette approche souffre de la limitation du nombre de traits utilisables en entrée, qui est de 80 dans notre cas. Les perspectives pour ce système seraient d'utiliser un plus grand nombre de trait avec leur score de saillance comme pondération.

Notre seconde approche semble quant-à elle prometteuse car donnant de meilleurs résultats que la précédente. De plus, une marge de progression important peut sens doute être réalisée en affinant le

filtrage des termes selon le corpus considéré. Cela permettrait notamment de limiter les pertes en précision due à une surreprésentation d'une classe dans la liste des traits retenus.

Dans tous les cas, il serait probablement bénéfique de pouvoir travailler au niveau de la phrase, plutôt que de considérer le document comme un tout. Ceci nous donnerait en sus la possibilité de rejeter les phrases de narration, comme celles contenus dans les synopsies de film, afin de ne pas comptabiliser des indices *positifs* ou *négatifs* n'indiquant pas de prise d'opinion de l'auteur.

Références

CRESTAN E. (2004). *Contextual semantics for WSD*, dans les Actes de Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, p. 101-104, Barcelone, Espagne.

GINI C. (1912). *"Variabilità e mutabilità"* Reimprimé dans Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).

GREFENSTETTE G., QU Y., SHANAHAN J. G. & EVANS D. A. (2001). *Coupling niche browsers and affect analysis for an opinion mining application.* Dans les Actes de RIAO-2004.

JOACHIMS T. (1997). *Text categorization with support vector machines: Learning with many relevant features*. Technical Report 23, Universitat Dortmund, LS VIII.

KULLBACK S. & LEIBLER R. A. (1951). *On information and sufficiency*, Annals of Mathematical Statistics 22: 79-86.

KUSHAL D., LAWRENCE S. & PENNOCK D. (2003). *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*. Dans les Actes de WWW2003, Budapest, Hungary, 20-24 Mai 2003, p 519–528.

RÜPING S. (2000). mySVM-Manual, University of Dortmund, Lehrstuhl Informatik 8, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.

SCHEIN A. I., POPESCUL A. & UNGAR L. H. (2002). *Methods and Metrics for Cold-Start Recommendations*. Dans les Actes de the XXV Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland.

SCHÜTZE H., HULL D. & PEDERSEN J. (1995). *A comparison of Classiers and document representations for the routing problem*. In International ACM SIGIR Conference on Research and Development in Information Retrieval.

WILSON T., WIEBE J. & HOFFMANN P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis.* Dans les Actes de Human Languages Technology Conference/EMNLP 2005.

YANG Y. & PEDERSEN J. (1997). *A comparative study on feature selection in text categorization*. In International Conference on Machine Learning (ICML).