

Défi: Classification de textes français subjectifs

Michel Génereux et Marina Santini

Natural Language Technology Group
University of Brighton, United Kingdom
{M.Genereux,M.Santini}@brighton.ac.uk

Résumé : Dans cet article, nous présentons le résultat de la classification de textes selon des critères subjectifs. La méthode proposée n'est pas nouvelle en soi, mais elle présente une brochette de traits et méthodes de normalisation des vecteurs de traits qui eux constituent une approche originale. Après une phase de réglage durant laquelle nous mettons au point une combinaison de traits et méthodes de normalisation susceptible de fournir les meilleurs résultats, nous soumettons les corpus de test à notre système. Les résultats obtenus, bien que modestes, nous permettent de tirer des conclusions intéressantes sur la validité et l'utilité d'une telle approche.

Mots-clés : Classification, Subjectivité, Traits, Normalisation

1 Introduction

La tâche demandée aux participants de DEFT 2007 était de classer des textes selon qu'ils ont un argumentaire plutôt *positif*, *négatif* ou *neutre*. Les textes proviennent de divers domaines : des critiques de films, livres, spectacles et bandes dessinées (corpus 1), des tests de jeux vidéo (corpus 2), des relectures d'articles de conférences (corpus 3) et des débats parlementaires (corpus 4). Les méthodes permises sont libres (supervisées, non supervisées), mais les ressources mises en oeuvre doivent se limiter aux corpus d'entraînement fournis par le comité DEFT 2007. Cet article décrit les techniques que nous avons utilisées pour la classification. L'article est organisé comme suit : la section 2 présente la méthode d'apprentissage automatique, les traits, le dictionnaire bilingue et les méthodes de normalisation utilisés. La section 3 est dédiée à la phase de mise au point alors que la section 4 présente les résultats de la tâche proprement dite. Nous discutons et concluons l'article aux sections 5 et 6 respectivement.

2 Méthodologie

2.1 Méthode d'apprentissage automatique

Nous avons arrêté notre choix sur la méthode d'apprentissage automatique dites *Support Vector Machine* (SVM). Cette méthode fût utilisée pour la première fois dans la classification de texte par (Joachims, 1997). Elle a fait ses preuves dans la classification de documents d'opinion, incluant le style (Diederich *et al.*, 2000), et elle a le grand avantage de pouvoir prendre en compte une grande quantité de traits, caractéristique essentielle en ce qui concerne notre approche. Durant la phase d'entraînement, l'algorithme construit un hyperplan qui sépare de façon maximale les exemples positifs et négatifs. La classification de nouveaux exemples consiste à trouver de quel côté du plan cet exemple se trouve. Cette méthode peut être adaptée pour plus de 2 classes. Le logiciel Weka¹, disponible gratuitement, fût utilisé.

2.2 Traits

On peut diviser notre répertoire de traits en 3 groupes : catégories grammaticales (*Adjectifs*, *Noms*, *Verbes* et *Adverbes*), facettes linguistiques fonctionnelles (*Facettes*) et groupes de termes à connotation émotive (*WordNet-Affect* et *Big-Six*) :

¹Le logiciel est disponible gratuitement à <http://www.cs.waikato.ac.nz/ml/weka/>. La méthode utilisée fût SMO avec les paramètres suivants : -C 1.0 -E 1.0 -G 0.01 -A 250007 -L 0.0010 -P 1.0E-12 -N 0 -M -V -1 -W 1

Groupe 1-Adjectifs, Noms , Verbes et Adverbes Ces catégories grammaticales ont la capacité d'exprimer une émotion ou un jugement subjectif (Turney, 2002).

Groupe 2-Facettes linguistiques fonctionnelles Dans la classification de documents selon leur *genre* (Santini, 2007), ces facettes ont donné de bons résultats. Elles sont données en Annexe A.

Groupe 3-Termes à connotation émotive Ces termes ont été classifiés par d'autres chercheurs comme ayant une composante émotive particulière. WordNet-Affect (Strapparava & Valitutti, 2004) est une extension affective de WordNet². Les termes sont divisés en *positifs*, *négatifs* et *neutres*³. Le groupe de *Big-Six* (Ekman, 1972) se base sur des études en psychologie et réorganise WordNet-Affect selon les six émotions de base suivantes : *colère*, *joie*, *tristesse*, *dégoût*, *peur* et *surprise*. Un extrait de ce groupe est donné en Annexe B.

Chaque terme appartenant à un des trois groupes et qui se qualifie comme trait se voit assigné une catégorie grammaticale à l'aide de Tree-Tagger⁴. Pour éviter de compter les négations (e.g. «Ce n'est pas un bon film.»), nous avons évité de comptabiliser tout terme (ici «bon») entre la particule négative *ne* et le délimiteur de fin de phrase.

2.3 Dictionnaire bilingue

En raison du manque de ressources en français pour certains calculs de normalisation nécessitant l'accès à un corpus ou lexique en anglais, certaines manipulations ont nécessité la création d'un dictionnaire bilingue anglais-français. Ce dictionnaire se compose de 1244 termes traduits manuellement et provenant des groupes de traits 2 et 3. Ce dictionnaire est nécessaire pour le calcul des facteurs de normalisation **pmi**, **sim**, **sent** et **vrai**, décrits dans la section suivante.

2.4 Méthodes de normalisation du nombre de traits

Chaque vecteur représentant un document attribue une valeur numérique à chacun des traits. La méthode de comptage la plus simple est dite *binnaire*, où seulement la présence (valeur 1) ou l'absence (valeur 0) est prise en compte. Une autre façon simple de comptage est la *fréquence*, où le nombre d'apparitions du trait dans le document est directement pris en compte, souvent normalisé à une longueur de document fixe (dans notre cas, 1000 mots). Nous avons considéré dans nos expériences d'autres façons de normaliser la *fréquence*, en multipliant celle-ci par l'un ou plusieurs des facteurs suivants :

idf De l'anglais *Inverse Document Frequency*. Permet d'évaluer l'importance d'un terme *i*, la supposition étant que l'importance d'un terme diminue à mesure qu'il apparaît dans une proportion grandissante de documents faisant partie du corpus. La formule exacte est :

$$idf_i = \log \frac{D}{d_i}$$

où *D* est le nombre total de documents et *d_i* est le nombre de documents dans lequel le terme *i* apparaît.

so-pmi-ir De l'anglais *Semantic Orientation - Pointwise Mutual Information - Information Retrieval*. Cette stratégie permet de calculer l'orientation sémantique (SO) de termes (textes) en calculant leur degré d'association (*A*) avec une liste de mots positifs et négatifs (*P* et *N*). Elle fût utilisée par (Turney, 2002) pour classer des termes selon leur niveau de *sentimentalité*, qui peut être plus ou moins négative ou positive. Cette mesure, appelée SO-A, peut s'exprimer mathématiquement de la façon suivante :

$$\sum_p^P A(\text{terme}, p) - \sum_n^N A(\text{terme}, n)$$

Notons que la quantité de termes *P* doit être égale à la quantité de *N*. Pour calculer SO-A, (Turney, 2002) a recouru à la notion de PMI-IR. PMI (Church & Hanks, 1989) entre deux termes est définie comme :

$$\log_2 \frac{\text{prob}(\text{terme}_1 \text{ est autour de } \text{terme}_2)}{\text{prob}(\text{terme}_1) * \text{prob}(\text{terme}_2)}$$

²<http://wordnet.princeton.edu/>

³Il y a aussi une liste de termes *ambigus*.

⁴<http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/>

PMI est positif quand deux termes ont tendance à apparaître ensemble et négatif quand ils sont en distribution complémentaire. PMI-IR est indicatif du fait qu'en IR, les occurrences multiples d'un terme dans un même document ne compte que pour une seule occurrence ; selon (Turney, 2002), cela semble fournir une meilleure mesure de SO, plus résistante au bruit. En calculant les probabilités à l'aide du nombre de documents (nd) extraites tel que fournie en IR, cela nous donne, pour PMI-IR :

$$\log_n \frac{D * (nd(terme_1 \text{ AUTOUR } terme_2) + 1/D)}{(nd(terme_1) + 1) * (nd(terme_2) + 1)}$$

où D est le nombre total de documents dans le corpus. Les termes de références positifs P employés furent l'équivalent anglais de : *bon, gentil, excellent, positif, chanceux, correcte, supérieur* et les termes de références négatifs N *mauvais, méchant, pauvre, négatif, malchanceux, fautif, inférieur*. Les valeurs d'aplanissement (*smoothing*) (1/D et 1) sont choisies pour que PMI-IR soit zéro pour les termes qui ne sont pas dans le corpus, un terme est considéré comme étant *AUTOUR* d'un autre terme s'il est à l'intérieur d'une fenêtre de 20 mots et \log_2 a été remplacé par \log_n , puisque le logarithme naturel est plus commun dans la littérature et que cela ne fait aucune différence pour l'algorithme. Nous avons utilisé le corpus de Waterloo⁵ contenant approximativement 46 millions de pages (documents). Chaque terme n'apparaissant pas dans le dictionnaire s'est vu attribué une valeur so-pmi-ir neutre (0).

so-sim Cette fois nous utilisons une mesure de similarité entre deux termes obtenue à l'aide de WordNet pour calculer SO-A. Cette approche est similaire à (Kamps & Marx, 2002), où la similarité est calculée en utilisant simplement un comptage des arcs séparant deux termes dans WordNet, une technique semblable au calcul effectué lorsqu'on veut connaître la relation génétique entre deux personnes à travers leurs ancêtres communs (Budanitsky & Hirst, 2001). Seuls les noms, verbes, adjectifs et adverbes peuvent avoir une affinité sémantique dans WordNet. Les termes de références positifs P employés furent l'équivalent anglais de : *bon-bonifier, gain-gagner, excellence-exceller, supériorité-surpasser* et les termes de références négatifs N *mauvais-empirer, perte-perdre, pauvreté-appauvrir, négationnier*. Chaque terme n'apparaissant pas dans le dictionnaire s'est vu attribué une valeur **so-sim** nulle. Le «package» Perl WordNet : :Similarity⁶ fût utilisé pour le calcul.

sen (Esuli & Sebastiani, 2006) fournit une ressource de valeur appelée *SentiWordNet* dans laquelle chaque *synset s* est associé à trois valeurs numériques décrivant le degré d'objectivité et de subjectivité (positif et négatif). La somme des trois valeurs doit être 1, ce qui veut dire que chaque terme peut posséder, à des degrés divers, plus d'une propriétés en même temps. Une mesure unique de subjectivité peut donc être obtenue pour chaque terme faisant partie de *SentiWordNet*. La méthode utilisée pour le développement de *SentiWordNet* est basée sur l'analyse quantitative des commentaires associés aux *synsets* en entraînant un ensemble de classeurs pour 3 classes (positif, négatif et objectif) (Esuli & Sebastiani, 2005). La valeur attribuée à chaque classe correspond à la proportion de classeurs qui ont choisi cette classe en particulier. *SentiWordnet* a été évalué favorablement à l'aide du *General Inquirer* (Stone *et al.*, 1966). Un extrait de *SentiWordnet* est fournie en annexe C. Chaque terme n'apparaissant pas dans le dictionnaire s'est vu attribué une valeur **sen** nulle.

hum Une liste de termes annotés manuellement comme étant soit positif (+1), soit négatif (-1) par (Turney, 2002). Un extrait de cette liste est fournie en annexe D. Chaque terme n'apparaissant pas dans le dictionnaire s'est vu attribué une valeur **hum** nulle.

binf Normalisation hybride, elle permet de faire une distinction entre le groupe de trait 1 (normalisé de façon *binnaire*) et les groupes de traits 2 et 3 (normalisés selon la *fréquence*).

3 Mise au point

Dans la phase de mise au point, nous avons tenté d'établir une combinaison de traits et de méthodes de normalisation qui soit le mieux adapté pour l'ensemble des tâches de classification. L'ensemble du corpus d'entraînement de *Relectures* nous a servi d'étalon (227 textes classés 0, 278 classés 1 et 376 classés 2). Nous avons conservé les 500 traits les plus fréquents en ce qui concernent les traits de catégories grammaticales (adjectifs, noms, verbes et adverbes). Nous avons établi 18 combinaisons arbitraires de traits et

⁵<http://canola1.uwaterloo.ca/>

⁶<http://www.d.umn.edu/~tpedersen/similarity.html>

de méthodes de normalisation. Ces combinaisons se nomment *I, 2, 3, 4, 5, 6, 7, A, B, C, D, E, F, G, H, I, J* et *K*. Le tableau 1 illustre ces différentes combinaisons. Par exemple, la combinaison *J* est formée des 500 adjectifs et adverbes les plus fréquents tels que comptabilisés dans le corpus d'entraînement, ainsi que la fréquence totale des trois catégories de WordNet-Affect. Nous avons soumis les 18 combinaisons

Traits/Normalisation	binaire	fréquence	idf	so-pmi-ir	so-sim	sen	vrai	binf
Adjectifs	1ABHI	2 FG	3G	4	5	6	7	J
Noms	1 B	2	3	4	5	6	7	
Verbes	1 B I	2	3	4	5	6	7	
Facettes	1	2C	3	4	5	6	7	
WordNet-Affect	1	2 D K	3	4	5	6	7	J
Big-Six	1	2 E	3	4	5	6	7	
Adverbes	HI							J

TAB. 1 – Mise au point : 18 combinaisons de traits et méthodes de normalisation

à l'algorithme de classification pour le classement des textes de *Relectures*, chaque fois en utilisant une validation croisée (10-fois). Les résultats peuvent être visualisés en ordre décroissant d'exactitude à l'aide du tableau 2. Puisque DEFT 2007 permet de soumettre jusqu'à 3 exécutions différentes, nous avons donc

Combinaison	H	J	A	I	1	F	3	2	B
Exactitude	50.6	50.3	49.4	47.2	45.6	45.3	44.7	44.4	43.7
Combinaison	E	G	4	D	6	5	K	7	C
Exactitude	42.9	42.7	42.7	42.7	42.2	42.2	42.1	41.7	41.4

TAB. 2 – Mise au point : classification des *Relectures* avec différentes combinaisons de traits et méthodes de normalisation

choisi d'utiliser les combinaisons H, J et A pour la classification des corpus de test. Pour plus de clarté, nous les répétons ici :

Expérience 1 - Combinaison H Traits : adjectifs et adverbes.

Normalisation : binaire.

Expérience 2 - Combinaison J Traits : adjectifs, adverbes et Wordnet-Affect.

Normalisation : binaire pour adjectifs et adverbes, fréquence pour Wordnet-Affect.

Expérience 3 - Combinaison A Traits : adjectifs.

Normalisation : binaire.

4 Expériences

Les tableaux 3, 4, 5 et 6 détaillent les résultats de 3 expériences (H, J et A) sur les 4 corpus. Pour des raisons de rapidité de traitement, nous avons fait les choix suivants : à l'exception du corpus 3 (500 traits), le nombre maximum de traits utilisés fût 100, alors que le corpus 4 s'est vu amputé de 80% (le nombre exacte de texte d'entraînement est indiqué entre parenthèses). Chaque tableau est divisé de telle sorte que les validations croisées de la phase d'entraînement sont d'abord présentés, suivis par les résultats sur les fichiers de test. Chaque ligne de la matrice de confusion indique la distribution des textes parmi les classes. Par exemple, dans le tableau 3, combinaison H, des 309 textes étiquetés *zéro*, seulement 76 ont été classés correctement.

5 Discussion

À l'exception d'un sous-groupe particulier du groupe 3 (WordNet-Affect), nos résultats de la phase de mise au point montrent que, dans le cadre d'une approche supervisée avec SVM, la meilleure façon d'obtenir des taux d'exactitude raisonnable est de s'en tenir aux traits familiers (adjectifs, adverbes) avec normalisation binaire. En soi ce résultat est intéressant, quoiqu'un peu surprenant dans le cas de traits du

Combi- naison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (309)	un (615)	deux (1150)	
H	56.2	0.425	0.246	0.311	76	61	172	zéro
		0.415	0.270	0.327	50	166	399	un
		0.618	0.803	0.699	53	173	924	deux
J	55.4	0.392	0.236	0.295	73	64	172	zéro
		0.399	0.259	0.314	55	159	401	un
		0.615	0.797	0.695	58	175	917	deux
A	55.7	0.382	0.220	0.279	68	34	207	zéro
		0.392	0.115	0.178	53	71	491	un
		0.593	0.884	0.710	57	76	1017	deux
Moyenne	55.8	0.470	0.429	0.423	Fin de l'entraînement			
H avec données de test		0.48	0.43	0.45	Tel que communiqué par DEFT 07			
J avec données de test		0.49	0.44	0.46	Tel que communiqué par DEFT 07			
A avec données de test		0.50	0.39	0.44	Tel que communiqué par DEFT 07			
Moyenne Test		0.49	0.42	0.45				
Tous les participants		0.53 ±0.10	0.48 ±0.07	0.50 ±0.07	Tel que communiqué par DEFT 07			

TAB. 3 – Corpus 1 (Critiques de films, livres, spectacles et bandes dessinées)

Combi- naison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (497)	un (1166)	deux (874)	
H	64.6	0.610	0.453	0.520	225	216	56	zéro
		0.639	0.703	0.699	113	820	233	un
		0.673	0.681	0.677	31	248	595	deux
J	62.1	0.568	0.435	0.493	216	224	57	zéro
		0.616	0.679	0.646	127	792	247	un
		0.651	0.649	0.650	37	270	567	deux
A	63.2	0.551	0.467	0.505	232	231	34	zéro
		0.617	0.669	0.642	160	780	226	un
		0.695	0.677	0.686	29	253	592	deux
Moyenne	63.3	0.624	0.604	0.613	Fin de l'entraînement			
H avec données de test		0.64	0.60	0.62	Tel que communiqué par DEFT 07			
J avec données de test		0.64	0.61	0.63	Tel que communiqué par DEFT 07			
A avec données de test		0.61	0.59	0.60	Tel que communiqué par DEFT 07			
Moyenne Test		0.63	0.60	0.62				
Tous les participants		0.69 ±0.10	0.64 ±0.09	0.66 ±0.09	Tel que communiqué par DEFT 07			

TAB. 4 – Corpus 2 (Tests de jeux vidéo)

type *Big-Six*, où l'on aurait pu s'attendre à mieux. Dans le cas des méthodes de normalisation, force est d'admettre que dans certain cas, la faible dimension de notre dictionnaire bilingue n'a probablement pas permis à ces facteurs de jouer un rôle déterminant. Notons aussi que l'utilisation de bigrammes comme traits constitue une voie de recherche intéressante. À ce stade, l'approche qui semble la plus prometteuse compte adjectifs et adverbes de façon binaire et un groupe de termes (WordNet-Affect) selon leur fréquence globale.

D'autre part, nous avons essayé un type de traits jamais utilisé auparavant dans la classification de textes d'opinion, les facettes linguistiques. Ces facettes, telles que présentées dans (Santini, 2007), sont des macro-traits qui peuvent être *interprétés fonctionnellement*. Par exemple, la facette *première personne* inclut les pronoms personnels singuliers et pluriels. Cette facette indique que le contexte de communication est relié à celui qui produit le texte. Une fréquence élevée de cette facette dans un texte signale une position impressionniste ou subjective. Alors que dans la plupart des tâches de classification de textes les traits sont utilisés individuellement sans plus d'interprétation, avec les facettes le but est d'interpréter une vue particulière

Combi- naison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (227)	un (278)	deux (376)	
H	50.2	0.435	0.445	0.440	101	70	56	zéro
		0.439	0.482	0.460	63	134	81	un
		0.602	0.551	0.575	68	101	207	deux
J	50.1	0.441	0.445	0.443	101	72	54	zéro
		0.436	0.482	0.458	59	134	85	un
		0.597	0.548	0.571	69	101	206	deux
A	47.9	0.427	0.436	0.431	99	76	52	zéro
		0.397	0.442	0.418	68	123	87	un
		0.590	0.532	0.559	65	111	200	deux
Moyenne	49.4	0.475	0.485	0.484	Fin de l'entraînement			
H avec données de test		0.47	0.47	0.47	Tel que communiqué par DEFT 07			
J avec données de test		0.46	0.46	0.46	Tel que communiqué par DEFT 07			
A avec données de test		0.43	0.44	0.43	Tel que communiqué par DEFT 07			
Moyenne Test		0.45	0.46	0.45				
Tous les participants		0.48	0.46	0.47	Tel que communiqué par DEFT 07			
		±0.05	±0.05	±0.05				

TAB. 5 – Corpus 3 (Relectures d'articles de conférences)

Combi- naison	Validation-Croisée 3-fois				Matrice de confusion		Classe
	Exactitude	Précision	Rappel	F-score	zéro (2080)	un (1380)	
H	62.7	0.643	0.852	0.733	1773	307	zéro
		0.564	0.288	0.381	983	397	un
J	63.4	0.650	0.849	0.736	1766	314	zéro
		0.576	0.309	0.403	953	427	un
A	64.6	0.651	0.886	0.751	1843	237	zéro
		0.623	0.284	0.390	988	392	un
Moyenne	63.6	0.618	0.578	0.566	Fin de l'entraînement		
H avec données de test		0.56	0.52	0.54	Tel que communiqué par DEFT 07		
J avec données de test		0.57	0.54	0.55	Tel que communiqué par DEFT 07		
A avec données de test		0.60	0.54	0.57	Tel que communiqué par DEFT 07		
Moyenne Test		0.58	0.53	0.55			
Tous les participants		0.65	0.63	0.64	Tel que communiqué par DEFT 07		
		±0.06	±0.06	±0.06			

TAB. 6 – Corpus 4 (Débats parlementaires)

dans la communication. Par exemple, il est pris pour acquis qu'une fréquence élevée de la facette *première personne* indique qu'on est en présence d'un texte ARGUMENTAIRE comme des COMMENTAIRES ou des OPINIONS. (Santini, 2007) s'est servi de 100 facettes, divisées en plusieurs sous-types (e.g. fonctionnelles, syntagmatiques ou HTML). Pour DEFT nous n'avons utilisé que 14 facettes. Quoique plusieurs des 100 facettes utilisées dans (Santini, 2007) étaient de natures grammaticales basées principalement sur la sortie d'un parseur-étiqueteur (Tapanainen & Järvinen, 1997) pour l'anglais, dans ce défi nous avons sélectionné un petit sous-ensemble (principalement lexical) exploratoire : les facettes sont montrées en annexe A. La classification sémantique de verbes⁷ en sept catégories est prise de (Biber *et al.*, 1999).

L'utilisation de facettes introduit deux innovations pour la classification de textes argumentaires. La première est reliée à la nature grammaticale des facettes. Alors que la plupart des recherches portant sur des textes affectifs (e.g. analyse de sentiments, classification d'opinion) sont basées sur des termes ayant une connotation affective (Hatzivassiloglou & Wiebe, 2000; Riloff & Wiebe, 2003), en l'occurrence des adjectifs et adverbes, les facettes mettent l'accent sur l'utilisation de signaux grammaticaux. Nous avons noté une variation dans l'utilisation de pronoms personnels à travers les différents types de textes. Par exemple, les textes de jeux vidéo réfèrent souvent directement aux joueurs en utilisant un pronom personnel de la

⁷Traduits en français pour ce défi.

deuxième personne, comme dans la phrase «Vous incarnez un guerrier [...]», alors que les débats parlementaires font souvent l'utilisation de pronoms personnels de la première personne qui mettent l'emphase sur la vue exprimées par l'auteur, comme dans «Nous avons passé des dizaines d'heures en juillet et en août [...] à analyser ce projet. Mon sentiment est qu'il répond bel et bien à l'évolution du monde [...]». De façon similaire, les verbes d'activité apparaissent plus souvent dans les textes sur les jeux vidéo, alors que les relectures d'articles se caractérisent plus par des verbes mentaux et de communication. Nous avons aussi fait l'hypothèse que les fréquences attribuées aux nominaux et prédicats (qui habituellement aident à faire la distinction entre l'écrit et le parlé) peuvent varier à travers les différents types de textes.

La deuxième innovation réfère à la nature composée des facettes ; en d'autres mots, les facettes sont des macro-traits, c'est-à-dire que chaque facette est composée d'un certain nombre de traits individuels qui partagent une interprétation sémantique et textuelle similaire. Nous avons défini les facettes comme des traits *interprétées fonctionnellement* parce qu'elles aident à l'interprétation et à la reconstruction du contexte de communication par l'intermédiaire de signaux linguistiques. L'utilisation de macro-traits comporte un avantage pratique. En fait, les facettes réduisent le risque d'*overfitting*, un phénomène qui apparaît habituellement quand un modèle statistique a trop d'attributs. Utilisées seules (expérience C), les facettes permettent de classer correctement 41% des relectures, ce qui nous semble encourageant, d'autant plus si on compare avec une expérience (A, exactitude 50%) où un grand nombre (500) de traits (adjectifs) ont été utilisés (voir table 2).

6 Conclusion

Dans ce défi nous avons adopté une approche classique supervisée (SVM avec traits reliés aux catégories grammaticales - groupe 1) pour la classification de textes à teneur subjective, l'intention de base étant d'améliorer les performances rapportées dans la littérature (pour l'anglais) en faisant appel à d'autres types de traits, des facettes linguistiques fonctionnelles (groupe 2) et une liste de termes à connotation émotive (groupe 3), en plus de facteurs de normalisation diverses.

Pour ce qui est des résultats du défi en soi, nous nous en tirons honorablement, avec un F-score à l'intérieur de l'écart-type autour de la moyenne, sauf pour les débats parlementaires (2 classes). Nous avançons l'hypothèse que la mise au point, concoctée sur un corpus avec 3 classes (Relectures), n'est pas optimale pour la classification avec 2 classes (Débats). Les F-scores varient entre 54% et 57% pour 2 classes (Débats) et entre 43% et 63% pour 3-classes. Dans tous les cas, ces résultats sont largement au-dessus de ce que l'on pouvait s'attendre en choisissant au hasard (33% et 25%). La validation-croisée que nous avons effectuée à l'interne fût une bonne prédiction des résultats du test. Le meilleur résultat est pour les jeux vidéo (63%). La matrice de confusion pour le corpus 1 (critiques) révèle un lourd penchant en faveur des classes d'entraînement les plus nombreuses, et nous suspectons qu'un corpus de relectures d'articles de conférences (3), avec son langage plutôt neutre, présente une difficulté particulière pour ce genre de classification (d'où le faible F-score de l'ensemble des participants), surtout dans le cas de systèmes basés exclusivement sur le contenu lexical comme le nôtre. D'autre part, nous nous sommes limités à un nombre restreint de traits (les 100 les plus fréquents dans le cas des corpus 1,2 et 4, 500 dans le cas du corpus 3), ainsi que d'un nombre réduit de textes (1/5 du corpus original) pour l'entraînement de la tâche 4.

En ce qui concerne aux traits linguistiques fonctionnelles, ces premiers résultats nous permettent d'être encouragé. En les combinant avec d'autres traits, par exemple dans l'expérience 1, l'exactitude atteint environ 46%, un résultat compétitif. Le but à long terme est d'augmenter le nombre de facettes utilisées et de trouver une combinaison idéale de facettes et de traits plus traditionnels dans l'espoir d'augmenter la performance générale.

Summary

In this article, we present the results of a text classification task according to subjective criteria. The proposed method is not new as such, but it is based on a set of features and normalizing methods for the feature vectors that do constitute an original approach. After a tuning phase in which we investigate which combinations of features and normalizing methods are the best, we submit the testing data to our system. The accuracy of our results are modest, but allow us to draw interesting conclusions on the validity and utility of such an approach.

Keywords : Classification, Subjectivity, Features, Normalization

Références

- BIBER, JOHANSSON, LEECH, CONRAD & FINEGAN (1999). *Longman Grammar of Spoken and Written English*. USA : Longman.
- BUDANITSKY A. & HIRST G. (2001). Semantic distance in wordnet : an experimental, application-oriented evaluation of five measures. In *NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.
- CHURCH K. W. & HANKS P. (1989). Word association norms, mutual information and lexicography. In *27th annual Conf. of the ACL*, p. 76–83 : New Brunswick, NJ :ACL.
- DIEDERICH J., KINDERMANN J., LEOPOLD E. & PAASS G. (2000). Authorship attribution with support vector machines.
- EKMAN P. (1972). Universal and cultural differences in facial expression of emotion. In J. COLE, Ed., *Nebraska Symposium on Motivation*, p. 207–282, Lincoln : University of Nebraska Press.
- ESULI A. & SEBASTIANI F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, Bremen, DE. Forthcoming.
- ESULI A. & SEBASTIANI F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining.
- HATZIVASSILOGLOU V. & WIEBE J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *International Conference on Computational Linguistics*.
- JOACHIMS T. (1997). *Text Categorization with Support Vector Machines : Learning with Many Relevant Features*. Rapport interne LS8-Report 23, Universität Dortmund. LS VIII-Report.
- KAMPS J. & MARX M. (2002). Words with attitude. In *1st International Conference on Global WordNet, Mysore, India*.
- RILOFF E. & WIEBE J. (2003). Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*. ACL SIGDAT.
- SANTINI M. (2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton.
- STONE P. J., DUNPHY D. C., SMITH M. S. & OGILVIE D. M. (1966). *The General Inquirer : A Computer Approach to Content Analysis*. MIT Press. MIT Press.
- STRAPPARAVA C. & VALITUTTI A. (2004). Wordnet-affect : an affective extension of wordnet. In *The 4th International Conference on Language Resources and Evaluation (LREC 2004)*, p. 1083–1086, Lisbon.
- TAPANAINEN P. & JÄRVINEN (1997). A non-projective dependency parser. In *The 5th Conference on Applied Natural Language Processing*, Washington, DC, USA.
- TURNER P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*.

A Quatorze Facettes Linguistiques Fonctionnelles

NOMINAL : noms, adjectifs et nombres
PRÉDICAT : verbes
PASSÉ : verbes au passé
PASSIF : verbes au passif
PREMIÈRE PERSONNE : je moi mon mien nous notre
DEUXIÈME PERSONNE : tien tu vous ton votre
TROISIÈME PERSONNE : il elle son eux lui
VERBES D'ACTIVITÉ : faire saisir aller donner prendre venir utiliser
quitter montrer essayer acheter travailler déplacer suivre mettre
payer amener rencontrer jouer courir tenir tourner envoyer asseoir
attendre marcher transporter perdre manger surveiller atteindre
ajouter produire fournir choisir porter ouvrir gagner attraper
passer secouer fixer vendre dépenser appliquer former obtenir
réduire arranger battre vérifier couvrir diviser rapporter étendre
réparer suspendre joindre étendre tirer recevoir répéter sauver
partager sourire lancer visiter accompagner acquérir avancer
comporter emprunter brûler nettoyer grimper combiner contrôler
défendre délivrer creuser affronter engager exercer élargir explorer
VERBES DE COMMUNICATION: dire raconter appeler demander écrire parler
énoncer remercier décrire réclamer offrir suggérer admettre annoncer
répondre argumenter renier discuter encourager expliquer exprimer
insister mentionner noter proposer publier citer répliquer reporter
taire signer chanter affirmer enseigner avertir accuser reconnaître
adresser aviser appeler assurer défier plaindre consulter convaincre
déclarer demander stresser excuser informer inviter persuader
téléphoner prier promettre questionner recommander remarquer répondre
spécifier jurer menacer presser accueillir chuchoter
VERBES MENTAUX: voir savoir penser trouver vouloir signifier nécessiter
sentir aimer entendre souvenir croire lire considérer supposer écouter
désirer demander comprendre attendre espérer supposer déterminer
consentir porter soucier choisir comparer décider découvrir douter
apprécier examiner confronter oublier haïr identifier imaginer soucier
apprendre occuper manquer noter planifier préférer prouver réaliser
réaliser rappeler reconnaître regarder souffrir souhaiter projeter
accepter permettre apprécier approuver évaluer blâmer soucier calculer
conclure célébrer confirmer compter oser mériter détecter écarter
distinguer expérimenter craindre pardonner deviner ignorer fier
interpréter juger justifier observer percevoir prédire prétendre
compter rappeler satisfaire résoudre étudier suspecter impressionner
VERBES CAUSATIFS: aider laisser forcer exiger affecter causer activer
assurer permettre prévenir assister garantir influencer permettre
VERBE D'OCCURENCE: devenir survenir changer mourir pousser développer
survenir survenir émerger tomber augmenter durer élever disparaître
couler briller couler glisser
VERBES EXISTENTIELS: devoir apparaître tenir rester vivre varier sonner
inclure impliquer contenir exister indiquer représenter tendre insérer
importer réfléchir associer rester révéler convenir mûrter concerner
constituer définir illustrer impliquer manque sembler devoir posséder
VERBES ASPECTUELS: débiter garder arrêter commencer continuer compléter
terminer finir cesser

B WordNet-Affect et Big-Six (extrait)

WordNet-Affect

 POSITIF joie rayonné exalté allègrement réjouir euphorisant triomphe
 NÉGATIF crainte effrayé terrible alarme impatient timide atroce
 NEUTRE apathie impassibilité rêveur langoureusement indifférence

Big-Six

 COLÈRE ombrage offense folie irritation enragement indignation outrage
 JOIE culte adoration chaleur triomphe unité sympathie tendresse
 TRISTESSE ennui poids apitoiement douleur oppression misère punition
 DÉGOÛT répugnance horreur nausée maladie révulsion revirement
 PEUR agitation effroi cercueil timidité suspens hésitation ombre
 SURPRISE admiration étonnement stupeur terreur merveille stupidité

C SentiWordNet (extrait)

TERME	SCORE POSITIF	SCORE NÉGATIF	SCORE NEUTRE
-----	-----	-----	-----
abandonné	0.000	0.000	1.000
accablant	0.250	0.250	0.500
adoration	0.625	0.000	0.375
affligé	0.125	0.625	0.250
affreux	0.000	0.625	0.375
agacer	0.000	0.625	0.375
agité	0.000	0.125	0.875
agressif	0.625	0.000	0.375
alarmé	0.000	0.500	0.500
aliéné	0.625	0.000	0.375
amoureux	0.500	0.125	0.375
apathique	0.000	0.750	0.250
appréhensif	0.000	0.625	0.375
approbation	0.500	0.125	0.375
ardent	0.375	0.125	0.500
atroce	0.000	0.500	0.500
attrister	0.000	0.750	0.250

D Termes classifiés manuellement (extrait)

NÉGATIF

abandon accablant affligé affligeant affreux aggravation aggraver
 agiter agressif agression agressivité alarme aliénation amertume
 anéantissement animosité antagonisme antipathie anxiété apathie
 apathique appréhensif appréhension atroce avare aversion belligérant

POSITIF

acclamation accomplir admiration admirer affection affectueux
 agréable ajustement alerte allégresse amical amour apaiser apprécier
 approbation approuver ardeur attachement avide bienfaisant
 bienveillance bienveillant bon bonheur bouillant calme captivation