TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet

Romain Vinot[1], Natalia Grabar[2,3], Mathieu Valette[2,4]
[1] Ecole Nationale Supérieure des Télécommunications
[2] Centre de Recherche en Ingénierie Multilingue - INALCO
[3] STIM - DIAM, AP-HP Pitié-Salpêtrière, Université Paris 6
[4] UMR 7114 CNRS/Paris X (MoDyCo)
romain.vinot@enst.fr, ngr@biomath.jussieu.fr, mathieu.valette@free.fr

Mots-clefs – Keywords

Classification automatique, Rocchio, kPPV, SVM, Internet, filtrage de l'information
Text classification, Rocchio, kNN, SVM, Internet, information filtering

Résumé - Abstract

Le filtrage de contenus illicites sur Internet est une problématique difficile qui est actuellement résolue par des approches à base de listes noires et de mots-clés. Les systèmes de classification textuelle par apprentissage automatique nécessitant peu d'interventions humaines, elles peuvent avantageusement remplacer ou compléter les méthodes précédentes pour faciliter les mises à jour. Ces techniques, traditionnellement utilisées avec des catégories définies par leur sujet (économie ou sport par exemple), sont fondées sur la présence ou l'absence de mots. Nous présentons une évaluation de ces techniques pour le filtrage de contenus racistes. Contrairement aux cas traditionnels, les documents ne doivent pas être catégorisés suivant leur sujet mais suivant le point de vue énoncé (raciste ou antiraciste). Nos résultats montrent que les classifieurs, essentiellement lexicaux, sont néanmoins bien adaptées : plus de 90% des documents sont correctement classés, voir même 99% si l'on accepte une classe de rejet (avec 20% d'exemples non classés).

Filtering of illicit contents on the Internet is a difficult issue which is currently solved with black lists and keywords. Machine-learning text categorization techniques needing little human intervention can replace or complete the previous methods to keep the filtering up-to-date easily. These echniques, usually used with topic classes (economy or sport for instance), are based on the presence or absence of words. We present an evaluation of these techniques for racism filtering. Unlike the traditional systems, documents are not categorized according to their main topic but according to the expressed point of view (racist or anti-racist). Our results show that these lexical techniques are well adapted : more than 90% of the documents are correctly classified, or even 99% if a rejection class is accepted (20% of the examples are not classified).

1 Introduction

L'expérience présentée ici a été faite dans le cadre du projet européen Princip (*Plateforme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet*)[1]. Princip s'inscrit dans l'action *Safer Internet Action Plan* de la Communauté européenne et concerne la détection de contenus illicites relatifs au racisme et au révisionnisme sur le Web.

Cette problématique, semblable à la détection d'autres contenus préjudiciables (pédophilie, trafic d'armes et de drogue, pornographie), peut être abordée avec deux approches : filtrage par liste noire et par mots clés. Les deux approches peuvent donner des résultats satisfaisants mais elles ont également des limites. Le *filtrage par liste noire* consiste à filtrer les pages dont l'adresse URL fait partie d'une liste constituée préalablement. Dans cette approche, il ne s'agit pas de la détection, mais uniquement du blocage de pages indésirables. La faiblesse principale ici est la mise à jour des listes noires : le Web évolue rapidement, de nouveaux sites apparaissent, d'autres se déplacent ou bien disparaissent. Le principe de *filtrage par mots clés* suppose que si un des mots clés révélateurs est trouvé dans une page HTML, l'accès à cette page doit être réglementé. Cette méthode s'avère faillible. Le sens des mots clés est susceptible de changer, à la fois en synchronie (en fonction du contexte) et en diachronie (selon les époques). Ainsi, *nègre* reçoit des valeurs radicalement différentes selon qu'il s'agit de l'*art nègre*, de la *tête de nègre*, du *nègre* d'un écrivain ou du *nègre* tel que le mot était couramment usité aux $XVIII^{eme}$ et XIX^{eme} siècles. En somme, l'insulte *sale nègre* n'est qu'un cas d'instanciation parmi d'autres. Avec des mots clés discrimininants, cette technique peut donc permettre de repérer les pages qui traitent du racisme mais ne permet pas de déterminer si les pages exposent un propos raciste ou non. Une autre limite est la variabilité du contenu et de marqueurs lexicaux dans les pages illicites. Les mots neutres à l'origine peuvent être récupérés et utilisés dans un contexte raciste (par exemple *jeune* sous-entend *jeune d'origine étrangère*). Une « réhabilitation » des mots (leur passage des contenus racistes vers des contenus neutres et antiracistes) est aussi possible. Les phénomènes linguistiques de créativité lexicale rendent également la tâche plus ardue. Par exemple, le verlan (*beurs*, *feujs*, *rabzas*), la modification de l'orthographe (*naigre* au lieu de *nègre*), les emprunts (*bougnoul*, *gnoul*, *Sieg Heil*). Tous ces exemples montrent que les mots clés doivent aussi être mis à jour régulièrement et de préférence en puisant les données directement à la source : dans les pages Web.

Dans notre projet, nous avons opté pour une utilisation de listes noires constituées grâce à une étude linguistique des documents. L'approche linguistique suppose que la combinaison d'indices venant de plusieurs niveaux d'unités linguistiques (caractères, morphèmes, catégories syntaxiques, expressions complexes, isotopies sémantiques, code HTML, etc.) et basée sur une analyse plus globale des documents Web permet de mieux cerner et profiler le contenu de ces documents. La détection et l'analyse d'indices de différents niveaux devient possible avec des outils lexicographiques, statistiques et de TAL.

L'expérience que nous présentons ici concerne l'application de systèmes de classification automatique (classifieurs) pour la détection de contenus racistes. Lorsqu'il est possible d'attribuer une ou plusieurs catégories à un ensemble de documents, ces systèmes peuvent, à partir d'un corpus où la catégorie de chaque document est connue a priori, « apprendre » une définition de ces catégories et ensuite les attribuer automatiquement à de nouveaux documents. Ces systèmes fonctionnent généralement au niveau lexical des documents et fondent leur décision sur

[1] Des informations plus amples peuvent être trouvées sur la page `http ://www.princip.net`.

la présence ou l'absence de mots.

Pour cette expérience, nous disposons d'un corpus catégorisé. L'algorithme de classification doit donc discriminer les pages comportant des propos racistes de celles comportant des propos antiracistes. Contrairement aux domaines d'application « traditionnels » de classifieurs (par exemple, sport ou économie) où une vue « ontologique » du domaine est possible avec des mots-clés précis et connus, la thématique que nous explorons ici est plus difficile à cerner a priori. D'autant plus que nous cherchons à distinguer des points de vue différents qui portent sur le même sujet. La question que nous nous posons alors est de voir jusqu'à quel point les techniques lexicales sont utilisables pour discriminer ce genre de différence. Autrement dit, existe-il des différences lexicales entre les discours racistes et antiracistes ?

Dans la suite de cet article, nous mentionnons des travaux sur l'utilisation de classifieurs textuels sur des tâches non purement thématique (section 2), ensuite nous décrivons les corpus d'apprentissage et de test (section 3) et les méthodes de traitement de ces corpus (section 4). Nous présentons ensuite les résultats et les discutons (section 5) et nous terminons avec une conclusion (section 6).

2 Travaux similaires

Les travaux autour du filtrage de courriers électroniques non sollicités (couramment appelé spams) à l'aide de classifieurs textuels exploitent différentes techniques d'apprentissage : Naïve Bayes et k plus proches voisins (Androutsopoulos *et al.*, 2000), algorithme génétique, boosting d'arbre de décision (Carreras & Márquez, 2001). Bien que la stricte définition de la catégorie spam ne soit pas thématique, il est possible de déterminer quelques thèmes généraux (pornographie, transfert de fonds, etc) facilement détectables par les indices lexicaux. Ces systèmes ont de très bonnes performances et commencent à être déployés en environnement industriel.

Le programme de recherche TDT (Topic Detection and Tracking) (TDT, 2001), sponsorisé par l'agence DARPA, concerne la classification d'un flux d'informations suivant l'évènement générateur (les nouvelles parlant de deux élections différentes doivent être classées dans deux catégories séparées car l'évènement sous-jacent n'est pas le même). Cette tâche est très difficile car elle est fondamentalement non thématique. La majorité des systèmes proposés utilisent néanmoins les classifieurs lexicaux standards et obtiennent des performances peu satisfaisantes.

Pang, Lee et Vaithyanathan (Pang *et al.*, 2002) ont utilisé des classifieurs thématiques standards pour discriminer des critiques de films positives et négatives. Peter Turney (Turney, 2002) utilise « l'orientation sémantique » (positive ou négative) des adjectifs pour déterminer la catégorie de critiques de plusieurs types d'objets (films, voitures, banques et destinations de voyages). Les deux évaluations ont des performances moyennes (bien meilleures qu'un classement aléatoire mais insuffisantes pour une utilisation réelle). Dans les deux cas, les auteurs expliquent les erreurs par le fait que tous ces systèmes fonctionnent par agrégation des indices trouvés sur chaque zone de texte. Or, le jugement final d'une critique n'est pas une simple somme des jugements de chaque sous-partie (un film peut avoir de bons acteurs et de bons dialogues et recevoir néanmoins une mauvaise appréciation globale).

Dans tous les cas, et comme dans celui du filtrage de propos racistes, il ne s'agit pas de déterminer le sujet principal du document : mail sollicité ou non, évènement générateur de la nouvelle, critique positive ou négative.

3 Constitution du corpus

Comme l'outil final de détection de contenus illicites est destiné à travailler sur le Web, les corpus de travail sont construits également à partir de données existant sur le Web. Nous utilisons les moteurs de recherche généraux que nous interrogeons avec des mots clés « sensibles ».

La constitution du corpus est faite en deux étapes : collecte massive de documents et ensuite leur catégorisation manuelle. La collecte de documents est faite de deux manières : interrogation manuelle et automatique (Grabar & Berland, 2001) de pages et de sites et leur rapatriement. Lors de la catégorisation manuelle, un document peut être catégorisé dans une des catégories prédéfinies : raciste, antiraciste, révisionniste, anti-révisionniste et non pertinent. En cas de doute, la catégorie indécidable est prévue. Les documents de cette catégorie sont analysés par un organisme compétent (Ligue Belge des Droits de l'Homme).

Au moment des expériences présentées ici, les corpus sont en cours de constitution. Le corpus que nous avons utilisé contient 739 documents dont 286 pages racistes, 444 611 occurrences, tirées de 43 sites et 453 pages antiracistes, 941 007 occ., tirées de 81 sites.

4 Description des algorithmes

Les algorithmes de classification fonctionnent au niveau lexical en prenant les tokens des documents comme unités descriptives (termes). Habituellement, les classifieurs sont utilisés sur des documents texte brut préalablement segmentés sur tous les caractères non-alphabétiques. Les unités linguistiques sont donc les mots, la ponctuation et les chiffres ayant été supprimés. Dans l'expérience que nous présentons ici, nous avons considéré les indices non textuelles (nombres, code HTML) comme des ancrages supplémentaires dans le texte et donc utiles pour la discrimination de contenus racistes et antiracistes. Nous avons donc effectué trois expériences : sur le texte brut et en conservant les nombres et le code HTML. Dans la suite de cette section, nous précisons la manière de traiter et de représenter les documents et décrivons les algorithmes de classification utilisés.

Représentation des documents Comme dans la majorité des algorithmes de classification, nous utilisons une représentation vectorielle (Salton *et al.*, 1975) des documents : le *sac de mots*. Ainsi chaque document d est représenté par un vecteur $[d]$ de R^n, où chaque coordonnée d_w est calculée par rapport à la fréquence $Occ(w, d)$ du terme w dans d selon la formule :

$$d_w = TFIDF(w, d) = log(1 + Occ(w, d)) * log(\frac{N}{N(w)})$$

où N est le nombre de documents du corpus et $N(w)$ est le nombre de documents dans lequel w apparaît au moins une fois. Un terme se voit donc attribuer un poids d'autant plus fort qu'il apparaît souvent dans le document et rarement dans le corpus complet. Chaque vecteur $[d]$ est ensuite normalisé en $[\underline{d}]$ afin de ne pas favoriser les documents les plus longs. Pour effectuer la normalisation, nous divisons chaque coordonnée d_w par la norme euclidienne du vecteur :

$$\underline{d_w} = \frac{d_w}{\sqrt{\sum_w d_w^2}}$$

Ces valeurs sont ensuite traitées pas les algorithmes de classification que nous utilisons : Rocchio, k-PPV et SVM.

Rocchio Rocchio (Rocchio, 1971) est un des plus vieux algorithmes de classification et l'un des plus simples. Un profil prototypique $[c]$ est calculé pour chaque classe c selon :

$$c_w = \frac{t}{N_c} \sum_{d \in c} \underline{d_w} - \frac{1-t}{N_{\overline{c}}} \sum_{d \notin c} \underline{d_w} \quad (1)$$

où N_c est le nombre de documents dans c, $N_{\overline{c}}$ est le nombre de documents n'appartenant pas à c, et t est un paramètre du modèle compris entre 0 et 1. Dans les situations où un document peut être attribué à une seule classe, t est souvent positionné à 1. Ces profils correspondent au barycentre des exemples (avec un coefficient positif pour les exemples de la classe et négatif pour les autres). Ces vecteurs sont également normalisés de la même façon que les documents. Le classement de nouveaux documents s'opère en calculant la distance euclidienne (équivalente au produit scalaire et à la similarité en cosinus puisque tous les vecteurs sont de norme 1) entre la représentation vectorielle du document et celle de chacune des classes ; le document est assigné à la classe la plus proche.

K plus proches voisins (k-PPV) k-PPV est un algorithme de la reconnaissance des formes qui a prouvé son efficacité face au traitement de données textuelles (Yang, 1997). La phase d'apprentissage consiste à stocker les exemples étiquettés. Le classement de nouveaux textes s'opère en calculant la distance euclidienne entre la représentation vectorielle du document et celles des exemples du corpus ; les k éléments les plus proches sont sélectionnés et le document est assigné à la classe majoritaire (le poids de chaque exemple dans le vote étant éventuellement pondéré par sa distance).

Support Vector Machine (SVM) SVM (Vapnik, 1995) est un des algorithmes les plus performants en classification textuelle (Joachims, 1998). L'idée principale est de trouver un hyperplan qui sépare au mieux les données et dont la séparation (ou *marge* : distance séparant la frontière du plus proche exemple) est aussi grande que possible. Cette recherche correspond à un problème d'optimisation au cours duquel des vecteurs supports (les exemples les plus proches de l'hyperplan) sont sélectionnés. L'hyperplan calculé permet ainsi de séparer l'espace en deux zones. Pour classer les nouveaux documents, on calcule dans quelle région de l'espace ils se situent et on leur attribue la classe correspondante.

5 Résultats et Discussion

5.1 Comparaison des performances des algorithmes

Pour mesurer les performances de classifieurs, nous utilisons la méthode standard de validation. Le corpus est aléatoirement divisé en deux parties : le corpus d'apprentissage avec lequel les classifieurs apprennent et le corpus de test avec lequel on calcule le taux de performance.

Les résultats obtenus sont présentés dans la partie gauche du tableau 1. Les performances relatives de chaque algorithme sont similaires à celles de (Yang, 1997) et (Joachims, 1998) : Rocchio est moins performant que les k-PPV, eux-mêmes étant légèrement inférieurs aux SVMs. Cette persistance des résultats tend à montrer que la nature du corpus traité n'est pas singulièrement différente des données traitées dans les problèmes classiques. Malgré le fait que la définition des classes ne soit pas thématique mais repose sur une analyse du discours, la description lexicale d'un document suffit à discriminer ces deux classes. Dans tous les cas, les performances sont assez impressionnantes, avec une moyenne supérieure à 0.90.

Algorithme	Performance
Rocchio	0.89
1-PPV	0.94
10-PPV	0.94
30-PPV	0.92
SVM	0.95

Algorithme	% d'exemples non classés	performance
10-PPV	0%	0.94
	10%	0.96
	20%	**0.99**
Rocchio	0%	0.89
	15%	0.93
	35%	0.98
SVM	0%	0.95
	10%	0.97
	20%	**0.99**

TAB. 1 – Performance des différents algorithmes

Tous les algorithmes présentés ici attribuent une valeur de confiance pour chaque prédiction (qui n'est pas représentée dans les résultats). Il est possible d'utiliser cette valeur afin d'affecter les exemples trop ambigus à une classe de rejet. Ainsi, les exemples dont la valeur de confiance est inférieure à un seuil prédéfini sont « rejetés ». Ce mode de fonctionnement est utile lorsqu'il est préférable d'avouer son ignorance plutôt que de faire une erreur. Avec cette classe de rejet, il est possible d'avoir plus de 99% d'exemples correctement classés en rejetant jusqu'à 20% des exemples (avec les K plus proches voisins ou les SVMs).

5.2 Analyse manuelle des résultats

À partir de l'analyse des 100 premières formes d'un *sac de mots* (les termes avec les poids les plus forts dans Rocchio) et des documents mal classés, nous allons tenter de déterminer quels sont les atouts et les limites de la classification algorithmique.

5.2.1 Les mots retenus par Rocchio

Le sac de mots raciste est constitué de mots qui participent à la construction des syntagmes identificatoires des sites du corpus. Ces items, qui ont les pondérations les plus fortes, dessinent la "signature lexicale" de chaque site. Du point de vue des documents, les items relèvent aussi bien du niveau textuel (slogans, mots d'ordre, qualifications des cibles) que péritextuel (sommaires, rubriques, titres), lesquels se répètent à l'identique dans plusieurs documents d'un même site. Ainsi, les mots *racaille* et *racailles* qui apparaissent parmi les quatre formes les plus déterminantes, constituent la dénomination euphémique privilégiée des cibles pour les auteurs du site `sos-racaille.org`.

De même, le mot *visage*, en 17ème position, participe à la construction d'une rubrique ("le vrai visage des potes") et du titre d'un article ("le vrai visage des islamistes")[2]. En fait, les mots du racisme qui ne sont pas spécifiques à un site particulier mais au discours raciste apparaissent au

[2] *Potes* est un moyen de qualifier les victimes particulièrement fréquent sur le site SOS-Racaille, lequel parodie celui de l'association antiraciste SOS-Racisme qui s'est fait connaître dans les années 80 par le slogan "Touche pas à mon pote".

sommet du sac de mots à condition qu'ils soient instanciés dans les sommaires (par exemple : *honte, envahie, agressions, désinformation*, etc.).

Quoi qu'il en soit, la forte prégnance des informations péritextuelles et identificatoires nous renseigne sur l'importance de la structure du document Internet pour les pondérations effectuées par les algorithmes de classification (en l'occurrence Rocchio). Ainsi, en l'état actuel de nos travaux, nous pouvons dire d'une part, que les algorithmes identifient et classent des documents (Internet) plutôt que des textes, pour autant que cette distinction soit pertinente, et d'autre part, qu'ils identifient ces documents en fonction de signatures lexicales plutôt que de modalités énonciatives racistes.

Le linguiste, qui est naturellement enclin à s'intéresser davantage au texte lui-même plutôt qu'au péritexte, peut s'étonner que les éléments qui lui semblent caractéristiques du discours raciste et non des sites racistes, soient relégués au second plan. Les résultats obtenus dessinent en effet très nettement une ontologie particulière (on pourrait dire "régionale") à quelques sites[3]. Or, le discours raciste, en tant qu'il procède d'une simple opposition (*nous* vs. *les autres*) avec dépréciation et péjoration des attributs des *autres*, ne relève pas à proprement parler d'une ontologie. Ainsi, on sait que le champ sémantique de la véracité (*vrai, véritable*, etc.), caractéristique du discours raciste (*nous* détenons la vérité ; *les autres* et leurs complices promeuvent le mensonge) apparaît être un critère de première importance. Or, il est absent des 100 premières formes du sac de mots analysé. Certes, à mesure que le corpus augmentera, des éléments moins spécifiques aux sites gagneront en importance. On peut également penser qu'une approche morphématique (mise en place dans le Projet Princip) neutraliserait les variations morpho-syntaxiques (par exemple *vraie, vrais*, etc.) que ne peuvent traiter nos algorithmes. Des analyses ultérieures devront en rendre compte.

Dans le sac de mots antiraciste de Rocchio, ce sont très nettement les éléments lexicaux appartenant à la rhétorique et aux modalités d'actions antiracistes qui apparaissent en premier lieu[4] : entités nommées (*mégret, vitrolles*), qualifications (*fascisme, extrême droite*), explications (*chômage, éducation*) et actions (*mouvement, associations, manifestation*). Les critères choisis par les algorithmes sont dans ce cas plus proches de ceux retenus par les linguistes.

5.2.2 Les erreurs de classifications

Du fait de l'orientation lexicale des techniques de classification algorithmiques, les écueils rencontrés sont sensiblement les mêmes que dans la détection par mot clés (voir sec. 1). Ainsi, il suffit que la connexion lexicale entre un texte antiraciste et le sous-corpus raciste d'apprentissage soit un peu trop élevé pour que ledit texte soit classé comme raciste. Mais l'inverse n'est pas vrai : les erreurs de classement des textes racistes ne relèvent pas à proprement parler des formes qui composent leur vocabulaire, mais des modalités d'expression. Autrement dit, si les algorithmes n'ont pas été capables de les classer correctement, c'est parce que le racisme y est policé et euphémisé.

D'une manière générale, il semble que les algorithmes aient mieux classé les documents antiracistes que les documents racistes. C'est, selon nous, l'indice que le discours antiraciste est

[3] Pour information, les trente premières formes du sac de mots raciste étudié sont : *cliquez, racailles, islamistes, racaille, antiblanc, photos, potes, moussaoui, sos, bois, amélie, site, brigadier, pitié, silence, avocats, blancs, visage, justice, dépêches, bannières, henri, écran, musulmans, désinformation, islamiste, attentats, terroristes, soutenir, tournantes*. Toutes ces formes sont actualisées dans le péritexte des documents.

[4] Précisons que dans les documents antiracistes, le péritexte est souvent moindre, voire inexistant.

d'une relative homogénéité, tandis que le discours raciste se manifeste de façons beaucoup plus variées, d'une part parce que, comme nous l'avons vu, il se situe en deçà de l'ontologie garante d'une unité lexicale, d'autre part, parce qu'il est actualisé dans différents discours et genres (discours idéologique, politique, genres pamphlétaire, essayiste, journalistique, etc.).

Les textes antiracistes mal classés sont essentiellement (1) des textes littéraires (paroles de chanson, extraits de romans en ligne), c'est-à-dire des textes qui ne répondent pas au style argumentatif caractéristique de l'antiracisme, et (2) des textes où, à des fins rhétoriques, les auteurs recourent abondamment à l'antiphrase et à la citation[5].

Les documents racistes mal classés, beaucoup plus nombreux, sont le plus souvent des textes politiques et idéologiques où le racisme n'est pas le thème principal et se trouve enchâssé dans une rhétorique de l'euphémisme (par exemple : éloge de personnalités vichystes, avec allusion à la situation contemporaine, article de webzine d'extrême droite s'en prenant ponctuellement aux populations immigrées, etc.). Les mesures statistiques sur lesquelles reposent les algorithmes de classification sont, en conséquence, inefficientes.

En résumé, on peut dire que devant deux ensembles de documents très différents quant à leur structure et aux modalités énonciatives, les algorithmes de classification et les linguistes adoptent une stratégie semblable en ce qui concerne le discours antiraciste, et différente avec le discours raciste : les algorithmes privilégient une approche globale du document et, d'une certaine façon, néglige la dimension énonciative, au sens classique, du discours raciste ; tandis que les linguistes délaissent le péritexte et se focalisent sur les modalités d'énonciation. Dans le cadre du Projet Princip, cette différence de "point de vue" nous a conduit à réévaluer le péritexte et à considérer le document Internet comme un objet textuel spécifique dont la complexité peut être étudiée de manière linguistique (nous avons ainsi attribués une dimension pragmatique au péritexte). - Un bel exemple de communication machine-homme.

5.3 Influence des nombres et du code HTML sur les résultats

Les résultats obtenus avec la prise en compte du code HTML et de nombres sont indiqués dans le tableau 2. Ces informations complémentaires influencent légèrement les résultats en améliorant les performances des algorithmes. L'influence est surtout visible avec la prise en compte du code HTML.

Algorithme	**Rocchio**	**1-PPV**	**10-PPV**	**30-PPV**	**SVM**
Sans chiffres ni HTML	0.89	0.94	0.94	0.92	0.95
Avec les chiffres	0.89	0.93	0.94	0.92	0.95
Avec source HTML	**0.94**	0.94	0.95	**0.96**	0.96

TAB. 2 – Performance avec des documents comportant les nombres et le code HTML

L'analyse des nombres et du code HTML discriminants apporte la confirmation à notre hypothèse que ces éléments constituent des ancrages supplémentaires dans les textes.

Nous constatons ainsi la présence de dates récentes (*2001*, *2002*) dans les documents racistes,

[5]Ainsi, k-PPV n'a pas été capable - et c'est bien compréhensible ! - de détecter l'ironie dans l'extrait suivant : "Ce n'est plus l'infection judéo-cosmopolite qui est dans la ligne de mire. Mais, l'invasion des allogènes qui, dans leur perversité, imposent leur loi, celle des bandes ethniques - On n'est plus chez nous, bordel !" (`http ://www.homme-moderne.org/kroniks/vlad/001001.html`)

très gourmands des faits divers qui, à travers des récits datés et documentés, permettent d'avoir un lien réel avec la vie quotienne. Les sites de ce type ont, en règle général, une mémoire courte.

En ce qui concerne les éléments du code HTML les plus discriminants, il s'agit de balises, d'attributs, de valeurs d'attributs et d'entités SGML. Ainsi, dans le corpus raciste, l'attribut HTML *pics* indique que les documents de ce corpus utilisent et affichent souvent les images (photos, dessins, caricatures, bannières, etc.). La balise *meta*, qui peut être utilisée pour noter une liste de mots clés ou d'autres informations, est également discriminante. Les polices de caractères *arial* et *verdana* semblent être dédié à ces documents, fait confirmé par une étude sur la présentation graphique des documents racistes (Nicinski, 2002).

Dans le corpus antiraciste, le terme *class* indique une utilisation plus fréquente de Java scripts dans ces documents. Par contre, avec les entités SGML (*eacute*, *egrave*, *ecirc*, *ocirc*, etc.) comme termes discriminants, il est difficile de savoir si leur présence discriminante n'est pas due à l'utilisation d'un éditeur HTML donné.

On en vient à se poser la question sur la représentativité et complétude de corpus étudiés. Dans quelle mesure les traits trouvés dans ces expériences sont discriminants de contenus racistes et antiracistes ? Vont-ils toujours être discriminants sur d'autres corpus et sur les documents du Web ? La convergence des résultats provenant de différentes études et expériences semblerait indiquer que certains de ces traits (mots et expressions, polices de catactères, utilisation des images et des balises *meta*, présence de dates récentes) sont révélateurs de contenus que nous cherchons à détecter. La traçabilité et l'explication d'autres traits est plus difficile et nuancée. Notons qu'une étude spécifique du code HTML est en cours, de même que l'étude de ces corpus avec d'autres approches et outils.

Il est clair que l'apprentissage effectué sur ces corpus devra être confronté à un corpus raciste plus grand, mais aussi à un corpus neutre ou bien le corpus composé de documents proches mais non pertinents. Il est clair aussi que ces corpus, en cours de constitution, doivent évoluer.

6 Conclusion

Nous avons présenté ici des expériences de traitement de corpus raciste et antiraciste avec des algorithmes de classification automatique. Ils fonctionnent sur un principe relativement similaire au filtrage par mots-clés mais permettent une réactualisation plus aisée puisque le réapprentissage est automatique dès lors que le corpus existe. Ces algorithmes, habituellement utilisés pour la classification thématique « classique », semblent être adaptés aux données « spéciales » des corpus traités. Leur performance est supérieure à 90% et même à 99% si on accepte une classe de rejet. Nous avons présenté des expériences faites avec du texte brut, mais aussi en tenant compte des nombres et du code HTML. Ces informations complémentaires améliorent très légèrement les performances.

A notre stade d'investigation, si l'on met de côté la question de l'euphémisation, la plus grande part des erreurs qui subsistent semblent être du même type que celles présentées dans les études de (Pang et al., 2002 ; Turney, 2002) : seule une petite partie du document est caractéristique du racisme. L'information est noyée dans le reste du texte et les algorithmes statistiques qui opèrent une moyenne de tous les termes du document ne parviennent pas à dégager la ou les phrases importantes. Cet échec semble lié à la prédominance du péritexte dans le choix des mots discriminants (cf. 5.1.1) : du fait de la redondance des sommaires et autres informations

propre à la structure des pages HTML, les documents *globalement* racistes sont mieux filtrés que ceux qui ne le sont que *localement*. Pour pallier ces faiblesses, dans l'outil final Princip, les documents seront également pris en charge par des modules plus fins avec utilisation d'informations plus complexes que la présence/absence de termes simples : chaîne de caractères (morphèmes), expressions complexes (lexies), isotopies, mesure de la cooccurrence, etc. Ces modules collaboreront selon un ensemble de stratégies prédéfinies.

Remerciements

Nous remercions François Rastier d'avoir noué cette collaboration. Nous remercions Monique Slodzian, François Rastier, François Yvon et Pierre Zweigenbaum pour leurs conseils, discussions et soutien précieux pour ce travail.

Références

ANDROUTSOPOULOS I., KOUTSIAS J., CHANDRINOS K. V. & SPYROPOULOS C. D. (2000). An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In N. J. BELKIN, P. INGWERSEN & M.-K. LEONG, Eds., *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, p. 160–167, Athens, GR : ACM Press, New York, US.

CARRERAS X. & MÁRQUEZ L. (2001). Boosting trees for anti-spam email filtering. In *Proceedings of RANLP-2001, 4th International Conference on Recent Advances in Natural Language Processing*.

GRABAR N. & BERLAND S. (2001). Construire un corpus web pour l'acquisition terminologique. In *Terminologie et intelligence artificielle*, p. 44–54, Nancy.

JOACHIMS T. (1998). Text categorization with support vector machines : Learning with many relevant features. In *ECML-98, Tenth European Conference on Machine Learning*, p. 137–142.

NICINSKI M. (2002). *Analyse et typologie des images dans les sites racistes*. Rapport interne, CRIM-INALCO. Mémoire de DESS, François Rastier (dir.).

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

ROCCHIO J. J. (1971). *The SMART Retrieval System : Experiments in Automatic Document Processing*, chapter 14, Relevance Feedback in Information Retrieval, p. 313–323. Gerard Salton (editor), Prentice-Hall Inc. : New Jersey.

SALTON G., WONG A. & YANG C. (1975). A vector space model for information retrieval. *Communications of the ACM*, **18**(11), 613–620.

TDT (2001). The Topic Detection and Tracking 2001, evaluation project. http ://www.nist.gov/TDT.

TURNEY P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 417–424, Philadelphia.

VAPNIK V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

YANG Y. (1997). *An evaluation of statistical approach to text categorization*. Rapport interne Technichal Report CMU-CS-97-127, Carnegie Mellon University.