Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films

Damien Poirier, Françoise Fessant, Cécile Bothorel, Emilie Guimier de Neef, Marc Boullé

Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films

Damien Poirier, Françoise Fessant, Cécile Bothorel
Émilie Guimier de Neef, Marc Boullé

France Telecom R&D, TECH / EASY
2 avenue Pierre Marzin, 22300 Lannion, FRANCE
prénom.nom@orange-ftgroup.com

Résumé. Les sites communautaires sont par nature des lieux consacrés à l'expression et au partage d'avis et d'opinions. *www.flixster.com* est un exemple de site participatif où se retrouvent chaque jour des dizaines de millions de fans dans le but de partager leurs impressions et sentiments sur les films. Une étude approfondie de cette richesse d'information permettrait une meilleure connaissance des utilisateurs, de leurs attentes, de leurs besoins. Pour y parvenir, une étape nécessaire est la classification automatique d'opinion. Dans ce papier nous décrivons trois approches permettant de classer des textes selon l'opinion qu'ils expriment. La première approche consiste à étiqueter les mots porteurs d'opinion à l'aide de techniques linguistiques, ces mots permettant par la suite de classer les textes. La deuxième approche est basée sur des techniques statistiques. La dernière approche est une approche hybride qui combine approche linguistique, pour prétraiter le corpus, et approche statistique, afin de classer les textes.

1 Introduction

Depuis l'émergence du Web 2.0 et des sites communautaires, une quantité croissante de textes non structurés prolifère sur la toile. Ces textes, généralement produits par les internautes, sont très souvent porteurs de sentiments et d'opinions sur des produits, des films, des musiques, etc. Ces données textuelles représentent potentiellement des sources d'information très riches permettant *a priori* de découvrir les attentes, désirs, besoins des utilisateurs ou encore de mesurer la popularité de certains produits, personnalités, décisions politiques, etc.

Le domaine de la fouille d'opinion peut-être divisé en trois sous-domaines (Pang et Lee, 2008) :

– l'identification des textes d'opinion, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisemment, on parle ici de classer les textes ou les parties de texte selon qu'il sont objectifs ou subjectifs ;
– le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions

contenant les opinions), chiffré (pourcentage, note), graphique (histogramme) ou encore imagé (thermomètre, étoiles, pouce levé ou baissé...) ;
– la classification d'opinion, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

Nous nous intéresserons ici uniquement à la classification d'opinion. Nous cherchons à déterminer les goûts cinématographiques d'utilisateurs à partir de l'analyse de leurs commentaires. Les données étudiées sont des textes d'opinion rédigés en anglais, chacun d'eux étant associé à un utilisateur et à un film mais également associé à une note attribuée par l'auteur au moment de la rédaction. Les textes de ce corpus présentent plusieurs particularités : ils sont en général très courts (une dizaine de mots) et le style dans lequel ils sont rédigés se rapproche du style utilisé dans les forums ou dans les systèmes de messagerie instantanée et comporte certaines abréviations de type SMS, des onomatopées, des fautes d'orthographes, des smileys, etc.

La problématique est donc de classer chaque commentaire selon l'opinion qu'il exprime une opinion positive ou une opinion négative. Pour ce faire, différents axes d'étude ont été envisagés :
– le premier axe consiste à se baser sur une approche linguistique. Le principe de ce type de méthodes consiste à construire, à l'aide d'outils de traitement automatique des langues, des lexiques de mots porteurs d'opinion et à classer les textes selon la présence ou l'absence de ces mots ;
– le deuxième axe concerne les méthodes statistiques, plus précisément les techniques d'apprentissage supervisé. Plusieurs de ces techniques existent et peuvent être appliquées au domaine de la classification d'opinion mais deux d'entre elles reviennent régulièrement et semblent fournir les meilleurs résultats. Il s'agit des Machines à Vecteur Support (SVM) et des Classifieurs Bayesiens Naïfs (NB). Nous nous sommes concentrés sur ces deux types de classifieurs
– le dernier axe combine traitements linguistiques et approche statistique. Nous avons appliqué des prétraitements linguistiques au corpus dans le but d'améliorer la représentation des textes, en amont de l'apprentissage supervisé.

Nous proposons un état de l'art avant de décrire les différentes approches testées.

2 Etat de l'art sur la classification d'opinion

L'analyse de texte en terme d'étude des sentiments, opinions ou points de vue n'est pas récente (Carbonell, 1979; Wilks et Bien, 1984). Cependant le domaine de la fouille d'opinion et de l'analyse des sentiments a pris une grande place dès le début des années 2000 avec l'arrivée du Web communautaire et la multiplication des forums sur la toile. Depuis ce jour, le domaine est devenu un enjeu majeur pour toute entreprise désireuse de mieux comprendre ce qui plait et déplait à ses clients ainsi que pour les clients qui souhaitent comparer les produits avant de les acquérir. Par exemple, Morinaga et al. (2002) expliquent comment ils vérifient les réputations de produits ciblés en analysant les critiques des clients. Ils recherchent tout d'abord les pages Web parlant du produit concerné et extraient les phrases qui expriment de

l'opinion. Ils classent ensuite les phrases selon qu'elles expriment une opinion négative ou une opinion positive et en déduisent la popularité du produit. Dans le même genre, Turney (2002) classe les commentaires selon deux catégories : *Recommended* et *Not Recommended*. Wilson et al. (2004) ajoutent à la classification selon la polarité, la force de l'opinion exprimée.

Enormément de travaux ont été effectués sur le sujet, et trois grandes catégories de méthodes peuvent être mises en avant : l'approche linguistique, l'approche statistique et l'approche hybride.

2.1 L'approche linguistique

La principale tâche dans cette approche est la conception de lexiques ou dictionnaires d'opinion. L'objectif de ces lexiques ou dictionnaires est de répertorier le plus de mots porteurs d'opinion possible. Ces mots permettent ensuite de classer les textes en deux (positif et négatif) ou trois catégories (positif, négatif et neutre).

Liu et al. (2005) décrivent un système, *Opinion Observer*, qui permet de comparer des produits concurrents en utilisant les commentaires écrits par les internautes. Ils ont une liste prédéfinie de termes désignant des caractéristiques de produits. Lorsqu'une de ces caractéristiques est présente dans un texte, le système extrait les adjectifs proches dans la phrase. Ces adjectifs sont ensuite comparés aux adjectifs présents dans leur dictionnaire d'opinion et ainsi, une polarité est attribuée à la caractéristique du produit.

Cette méthode nécessite donc la construction d'un dictionnaire d'opinion. Pour construire un tel dictionnaire, trois genres de techniques sont possibles :
- la méthode manuelle ;
- la méthode basée sur les corpus ;
- la méthode basée sur les dictionnaires.

La méthode manuelle demande un effort important en terme de temps mais il faut savoir que toutes les autres méthodes nécessitent également de créer initialement, de façon manuelle, un ensemble de mots et expressions porteurs d'opinions. Cet ensemble de mots est appelé *graine*. Il est ensuite utilisé afin de trouver d'autres mots et expressions porteurs d'opinions.

Une solution afin d'agrémenter cet ensemble de mots est donc l'utilisation de corpus de textes. Turney (2002) propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions déjà classés. Un mot apparaissant plus souvent à côté de mots positifs sera donc classé dans la catégorie positif et inversement. Yu et Hatzivassiloglou (2003) proposent une méthode similaire, mise à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé. D'autres méthodes (Pereira et al., 1993; Lin, 1998) utilisent également cette hypothèse dans le but d'agrémenter les lexiques d'opinion : deux mots ou groupes de mots ayant un fort degré d'apparition commune possèdent une forte proximité sémantique.

Une autre méthode basée sur le corpus permettant d'agrémenter le dictionnaire d'opinion consiste à utiliser les conjonctions de coordination présentes entre un mot déjà classé et un mot non classé (Hatzivassiloglou et McKeown, 1997; Kanayama et Nasukawa, 2006; Ding et Liu, 2007). Par exemple, si la conjonction *AND* sépare un mot classé positif dans le dictionnaire d'opinion et un mot non classé, alors le mot non classé sera considéré comme étant positif. À l'inverse, si la conjonction *BUT* sépare un mot classé positif et un mot non classé, alors le mot non classé sera considéré comme étant négatif. Les conjonctions utilisées sont les suivantes : *AND, OR, BUT, EITHER-OR*, et *NEITHER-NOR*.

La méthode basée sur les dictionnaires consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet (Miller et al., 1990). Afin de déterminer l'orientation sémantique de nouveaux mots, Hu et Liu (2004a) utilisent ces dictionnaires afin de prédire l'orientation sémantique des adjectifs. Dans WordNet, les mots sont organisés sous forme d'arbres (voir figure 1). Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation sémantique connue.

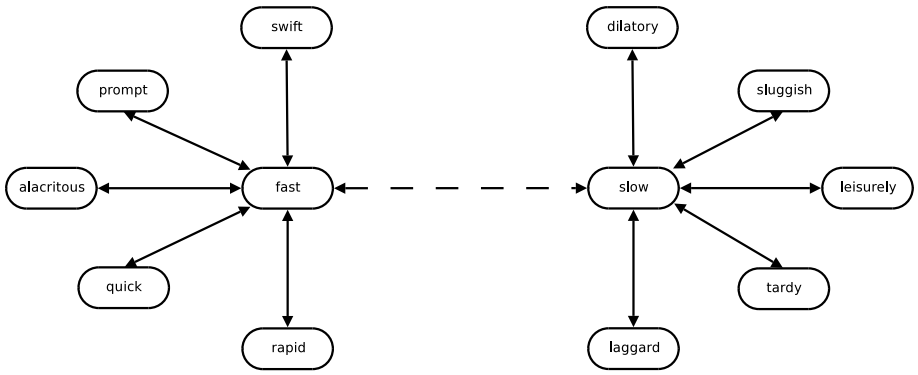

FIG. 1 – *Exemple d'arbre de synonymes et antonymes présents dans WordNet (flèche pleine = synonymes, flèche hachurée = antonymes)*

Afin de mesurer plus précisemment la force de l'opinion exprimée, un moyen utilisé est l'extraction des adverbes associés aux adjectifs. Pour ce faire, Benamara et al. (2007) proposent une classification des adverbes en cinq catégories : les adverbes d'affirmation, les adverbes de doute, les adverbes de faible intensité, les adverbes de forte intensité et les adverbes de négation et minimiseurs. Un système d'attribution de points en fonction de la catégorie de l'adverbe permet de calculer la force exprimée par le couple adverbe-adjectif.

Toutes ces catégories d'adverbes ne sont pas toujours prises en compte car elles n'ont pas la même importance au niveau de la prédiction de note. Les négations paraissent logiquement

être des termes importants à détecter, en plus des adjectifs et des verbes, car ils permettent d'inverser la polarité d'une phrase. Das et Chen (2001) proposent par exemple d'ajouter des mots dans le dictionnaire d'opinion comme "*like-NOT*" qui sont utilisés lors de la détection d'un couple like-négation. Les négations peuvent être *not, don't, didn't, never, ever...*

Le problème de la détection de la négation reste un problème très ouvert, les méthodes existantes n'étant pas réellement convaincantes. Ceci est aussi dû aux différentes façons d'utiliser la négation comme le sarcasme ou l'ironie.

Pour finir, il s'agit ensuite de déterminer la polarité d'une phrase à l'aide de ces dictionnaires. La solution la plus simple consiste à compter le nombre de mots positifs et le nombre de mots négatifs présents. S'il y a une majorité de termes positifs, la phrase est déclarée positive. À l'inverse, si les mots négatifs sont les plus nombreux, la phrase est déclarée négative. Les phrases possédant autant de mots négatifs que de mots positifs peuvent être déclarées neutres (Yu et Hatzivassiloglou, 2003), ou encore, la polarité de la phrase peut dépendre du dernier mot d'opinion parcouru (Hu et Liu, 2004a). On peut encore extraire plusieurs opinions dans une même phrase et les associer aux caractéristiques discutées (Hu et Liu, 2004b).

2.2 L'approche statistique

Les méthodes statistiques les plus utilisées sont les méthodes à apprentissage supervisé. Ce type de méthode consiste à représenter chaque commentaire comme un ensemble de variables, puis à construire un modèle à partir d'exemples de textes dont on connaît déjà le label. Le modèle est ensuite utilisé pour attribuer sa classe à un nouveau commentaire non étiqueté.

Pang et al. (2002) montrent que des techniques d'apprentissage automatiques offrent de meilleurs résultats que les méthodes linguistiques décrites précédemment. Ils précisent toutefois que les dictionnaires d'opinion utilisés ne sont peut-être pas optimaux. Pour faire leurs comparaisons, ils ont basé leurs expérimentations sur trois méthodes de classification automatique : un classifieur naïf bayesien, un algorithme de Machines à Vecteurs Support et un classifieur basé sur le principe d'entropie maximale.

Mais en fait, peu de travaux sont basés uniquement sur des méthodes statistiques. Le plus souvent, des prétraitements linguistiques sont effectués sur les textes, soit pour réduire le nombre de variables, ou encore pour sélectionner uniquement les traits grammaticaux susceptibles d'exprimer une opinion et ainsi éviter le bruit avec des mots inutiles pour ce type de classification. Ces approches constituent les approches dites hybrides.

2.3 L'approche hybride

Une première façon de faire est d'utiliser les outils linguistiques afin de préparer le corpus avant de classer les textes à l'aide d'outils d'apprentissage supersisé. Wilson et al. (2004) préparent les données à l'aide d'outils de Traitement Automatique des Langues afin de sélectionner un vocabulaire d'opinion. Ces mots pré-sélectionnés sont ensuite utilisés comme vecteurs de représentation des textes pour les outils d'apprentissage supervisé. Trois algorithmes d'apprentissage sont comparés : BoostTexter (Schapire et Singer, 2000), Ripper (Cohen, 1996)

et SVM^{light} (Joachims, 1999a). Nigam et Hurst (2006) utilisent des techniques provenant du Traitement Automatique des Langues afin de détecter dans les textes les mots et expressions porteurs d'opinion et ajoutent des marques dans le texte (traits grammaticaux et + ou - pour opinion positive et opinion négative). Ils utilisent ensuite l'apprentissage automatique pour classer les textes selon leur opinion générale.

Une autre façon de combiner les méthodes est d'utiliser les techniques d'apprentissage automatique dans le but de construire les dictionnaires d'opinion nécessaires à l'approche linguistique. Hatzivassiloglou et McKeown (1997) présentent une méthode ayant pour objectif de définir l'orientation sémantique des adjectifs pour la construction du dictionnaire d'opinion. Ils extraient tout d'abord tous les adjectifs du corpus à l'aide d'un analyseur syntaxique, puis utilisent un algorithme de clustering afin de classer les adjectifs selon leur polarité. Riloff et Wiebe (2003) combinent les deux approches afin de répertorier les expressions porteuses d'opinion qui, selon eux, sont plus riches que des mots pris individuellements. Turney et Littman (2003) utilisent une approche statistique pour classer un plus grand nombre de types de mots selon leur polarité : adjectifs, verbes, noms...

Une dernière façon d'utiliser conjointement les approches linguistiques et statistiques est de construire plusieurs types de classifieurs et de combiner leurs résultats, soit par des sytèmes de vote, soit par un algorithme d'apprentissage (Dziczkowski et Wegrzyn-Wolska, 2008).

3 Nos différentes approches

Nous présentons ici plusieurs techniques ayant pour objectif de classer des textes selon l'opinion qu'ils expriment. La première méthode est une approche linguistique et la deuxième, une approche statistique. Nous abordons également une approche hybride consistant à préparer et nettoyer le corpus à l'aide de méthodes linguistiques puis à classer les documents à l'aide d'outils d'apprentissage automatique.

Nous allons tout d'abord présenter le corpus sur lequel nous travaillons puis nous présenterons les différentes expérimentations réalisées.

3.1 Description du corpus

Le corpus d'expérimentations est extrait du site Flixster [1]. Ce site est un espace communautaire américain destiné aux amateurs de cinéma qui permet entre autres choses aux utilisateurs de se créer un espace personnel et de partager leurs impressions sur des films et des acteurs, le plus souvent en anglais. Les commentaires faits sur les films sont associés à une note comprise entre 0,5 et 5 précisant l'impression générale portée sur le film en question.

La principale difficulté de ce corpus est la grande variété de commentaires. En effet, que ce soit au niveau du style d'écriture ou de leur taille, les commentaires peuvent présenter de grandes disimilarités. Ceci rend la classification d'opinion plus difficile, parfois même pour

1. www.flixster.com

un humain. De plus, une grande partie du corpus est composée de messages plus proches des messages de forums que des critiques faites par des journalistes ou des professionnels. Ils présentent des caractéristiques telles que des smileys (" :-) "), des accumulations de ponctuation (" ! ! ! "), du langage SMS (" ur "," gr8 ") ou encore des étirements de mots (" veryyyyyy cooooool "). Le tableau 1 contient des exemples de commentaires associés aux notes.

Note	Commentaire
POS	Great movie !
NEG	this wasn't really scary at all i liked it but just wasn't scary...
POS	I loved it it was awsome !
NEG	I didn't like how they cursed in it.....and this is suppose to be for little kids....
NEG	Sad ending really gay
POS	sooo awsome ! ! (he's soo hot)
POS	This is my future husband lol (orlando bloom)
NEG	Will Smith punches an alien in the face, wtf ! ! ? ?
NEG	i think this is one of those movies you either love or hate, i hated it ! :o)

TAB. 1 – *Exemples de commentaires accompagnés de leur note.*

Le corpus extrait est composé de 60 000 commentaires. Comme nous l'avions annoncé, la taille des commentaires est très variable, allant de 1 à 518 mots, avec une moyenne de 13 mots par texte. La figure 2 donne une idée de la disparité rencontrée au niveau de la taille des différents commentaires. Par exemple, on observe que 60% des commentaires contiennent moins de dix mots.

Pour faciliter l'interprétation des résultats de classification, nous avons décidé de réduire l'espace des notes (10 classes allant de 0,5 à 5) à deux classes : positive (notes supérieures ou égales à 3) et négative (notes inférieures à 3). Nous expliquons dans la section 3.3.4 comment ce découpage a été déterminé. La moitié des commentaires composant le corpus appartient à la classe positive et l'autre moitié à la classe négative. Nous avons partagé le corpus en deux sous-ensembles de commentaires. Une partie pour les phases d'apprentissage (20 000 commentaires négatifs et 20 000 positifs) et une partie destinée aux tests (10 000 commentaires négatifs et 10 000 positifs). Tous les résultats qui suivent ont été obtenus à partir du même corpus.

Les seuls prétraitements appliqués sur le corpus, et qui sont valables pour tous les résultats qui vont suivre, sont la minusculisation de tous les caractères ainsi que la suppression de la ponctuation. Dans le cas de l'approche statistique, nous avons également supprimé tous les mots n'apparaissant qu'une seule fois dans le corpus dédié à l'apprentissage. Cela réduit de moitié l'espace de représentation des textes, et a un impact négligeable sur les résultats de la classification. Nous avons ensuite appliqué d'autres prétraitements linguistiques suivant l'expérimentation visée.

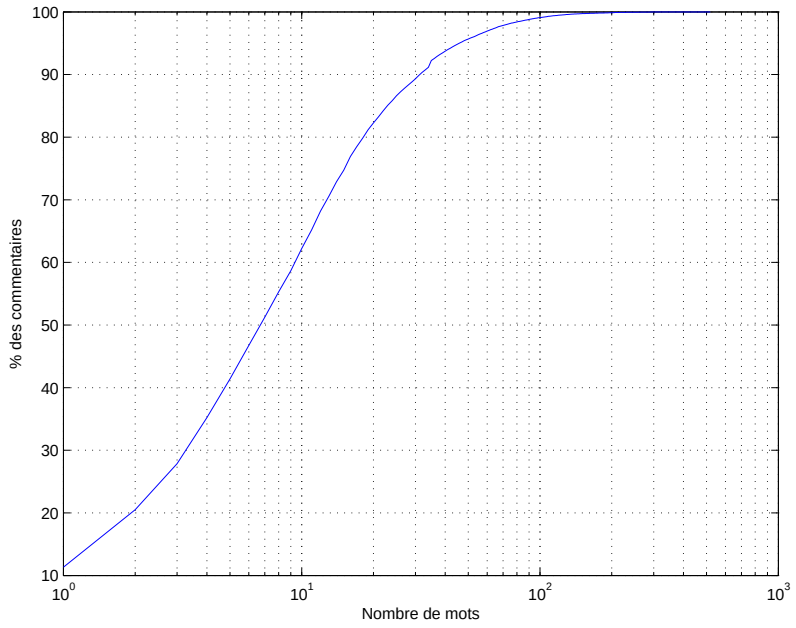

FIG. 2 – *Distribution cumulée du pourcentage de commentaires en fonction de leur taille.*

3.2 L'approche linguistique

Comme nous l'avons vu dans l'état de l'art, la première étape consiste à construire des dictionnaires contenant les mots porteurs d'opinion. Nous supposons donc que les opinions sont exprimées par certaines catégories de mots et que ces mots à eux seuls permettent de déterminer la polarité d'un texte. Nous allons tout d'abord expliquer comment nous avons construit nos dictionnaires puis nous verrons comment déterminer la polarité des textes et les résultats obtenus avec cette méthode.

3.2.1 Construction des lexiques

Nous avons fait le choix de construire deux lexiques distincts. Le premier d'entre eux contient tous les mots porteurs d'opinion positive et le second tous les mots porteurs d'opinion négative. Pour trouver les mots exprimant une opinion et les classer, nous avons tout d'abord séparé le corpus d'apprentissage en plusieurs parties en fonction des notes attribuées à chaque commentaire. Nous avons donc obtenu dix sous ensembles de commentaires notés respectivement de 0,5 à 5. Pour commencer nous avons appliqué, sur chacun des dix sous corpus, un analyseur syntaxique (Guimier de Neef et al., 2002) afin de lemmatiser et étiqueter chaque mot du texte. Nous nous sommes basés sur l'hypothèse que les adjectifs et les verbes étaient les deux traits grammaticaux les plus utilisés pour exprimer des opinions. Nous avons donc filtré les mots selon leur trait grammatical et leur fréquence dans chaque sous corpus, et conservé les adjectifs et les verbes ayant le plus d'occurences. Les mots séléctionnés apparaissant souvent dans les ensembles de commentaires notés de 4 à 5 ont été integrés dans le lexique de mots

positifs et inversement avec les mots apparaissant dans les commentaires notés de 0,5 à 2. Les lexiques ont ensuite été nettoyés manuellement afin de supprimer les termes n'exprimant *a priori* aucune opinion, ou encore les termes ambigus. Par exemple, le mot "*terrible*" n'apparait dans aucun des lexiques car il peut exprimer les deux types d'opinion.

Nous avons fait le choix de construire les dictionnaires d'opinion manuellement pour qu'ils ne contiennent que des mots vraiment spécifiques au corpus étudié. Nous pensons en effet que les lexiques d'opinion construits à l'aide des méthodes basées sur les dictionnaires (tels que WordNet), où l'on détermine la polarité des mots en fonction de leur synonymie, sont un peu trop aléatoires car beaucoup de mots peuvent avoir plusieurs sens selon le contexte. Nous avons aussi jugé que le corpus étudié n'est pas adapté aux constructions de lexiques d'opinion basées sur les corpus, les commentaires étant en règle générale très courts.

Au final, 183 mots *a priori* porteurs d'opinion ont été classés dans deux catégories. Le lexique de mots positifs contient 115 éléments et le lexique de mots négatifs en contient 68. Le tableau 2 présente des exemples de termes contenus dans les deux lexiques.

Mots postifs	good, great, funny, awesome, cool, brilliant, hilarious, favourite, well, hot, excellent, beautiful, cute, sweet ...
Mots négatifs	bad, stupid, fake, wrong, poor, ugly, silly, suck, atrocious, abominable, awful, lamentable, crappy, incompetent ...

TAB. 2 – *Exemples de mots contenus dans les lexiques d'opinion.*

3.2.2 Classification d'opinion

Cette dernière étape consiste à compter les mots porteurs d'opinion répertoriés dans les deux lexiques afin de déterminer la polarité de chaque commentaire. Pour ce faire, nous avons appliqué, sur le corpus de test, les mêmes prétraitements que précedemment, à savoir la suppression de la ponctuation et la lemmatisation, et nous avons conservé uniquement les adjectifs et les verbes. Le fait de ne garder que ces deux catégories de mots permet d'éviter quelques mauvaises interprétations de mots à double sens comme par exemple avec le terme "*like*" qui peut avoir plusieurs significations selon le contexte. Nous ne nous sommes pas essayé à des analyses linguistiques plus sophistiquées comme les analyses de structures grammaticales ou de dépendances au vu du style de langage utilisé dans le corpus.

La classification des commentaires se fait donc en comptant le nombre de mots d'opinion présents dans chaque commentaire. On calcul un score pour chaque commentaire en ajoutant 1 lorsque l'on rencontre un mot postitif, et en soustrayant 1 lorsque le terme rencontré est

négatif. Les commentaires possèdant donc une majorité de mots positifs (score positif) sont classés dans la catégorie *Commentaires Positifs* et inversement. Les commentaires possédant autant de mots positifs et négatifs sont ignorés. Il en est de même pour les commentaires qui ne contiennent aucun mot appartenant aux lexiques.

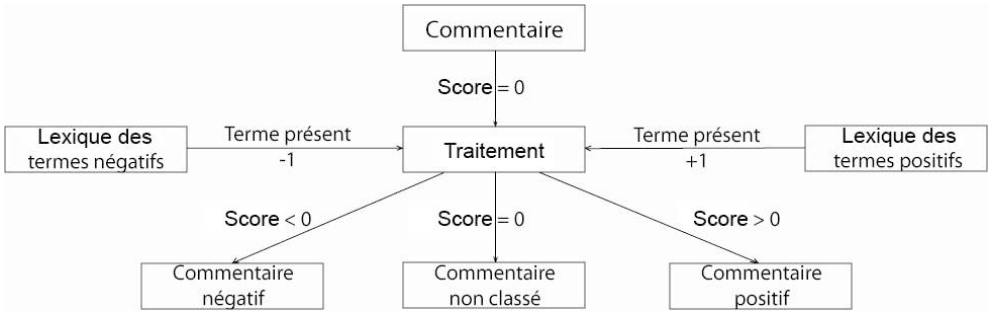

FIG. 3 – *Méthode de classification d'opinion.*

3.2.3 Résultats

Cette méthode nous a permis de classer 74% des 20 000 commentaires présents dans le corpus de test. Afin de pouvoir comparer les résultats obtenus avec les résultats des approches statistiques, nous avons décidé que tous les commentaires que la méthode n'a pas réussi à classer sont des commentaires négatifs. En d'autres termes, tous les commentaires qui ne sont pas positifs sont négatifs.

Dans le but de comparer ces résultats avec les résultats des autres méthodes, nous calculons quatre valeurs : l'accuracy, la précision, le rappel et le F_{score}. Les résultats de cette première expérimentation sont présentés dans la première colonne du tableau 3.

	Avec notre dictionnaire	Avec *General Inquirer*	Avec détection de la négation
Accuracy	72	66,4	69,6
Précision	72,8	65	70
Rappel	72,1	66,5	69,7
F_{score}	71,8	65,5	69,9

TAB. 3 – *Résultats des expérimentations de l'approche linguistique.*

Nous pouvons voir dans la matrice de confusion (tableau 4) que la grande difficulté de la tâche se situe au niveau de la classification des commentaires négatifs. Ce problème pourrait être dû aux lexiques que nous avons créés. En effet, le lexique de mots positifs contient pratiquement le double d'éléments en comparaison au lexique de mots négatifs. Mais le problème

n'est pas seulement la détection des commentaires négatifs mais également leur interprétation. En effet, plus des 2/3 des commentaires classés négatifs sont mal classés.

	Commentaires Positifs	Commentaires Négatifs
Commentaires Positifs Prédits	8 089	3 682
Commentaires Négatifs Prédits	1 911	6 318

TAB. 4 – *Matrice de confusion obtenue à partir de nos lexiques.*

Afin de vérifier la qualité de nos lexiques d'opinion, nous avons fait la même expérimentation en utilisant le lexique *General Inquirer* déjà construit par Stone et al. (1966) et Kelly et Stone (1975). Ce lexique contient 4 210 mots porteurs d'opinion (2 293 mots négatifs et 1 914 mots positifs).

L'utilisation de ce nouvel ensemble de mots d'opinion permet de classer plus de commentaires (78% contre 74% précédemment) mais les résultats de la classification sont moins bons (voir deuxième colonne du tableau 3). Ces scores peuvent être expliqués par le fait que le dictionnaire *General Inquirer* a été construit pour analyser des corpus traitant de sujet généraux, et non juste des commentaires de films. Nous pouvons également observer, grâce à la matrice de confusion (tableau 5), que le problème concernant la classification des commentaires négatifs est présent cette fois encore, bien que le lexique contienne plus de mots négatifs que de mots positifs. L'explication la plus plausible à ce problème paraît donc être l'utilisation de la négation qui est plus présente dans les commentaires négatifs que dans les commentaires positifs. Cette hypothèse se vérifie grâce aux résultats de l'approche statistique que nous allons présenter ensuite (voir section 3.3.6).

	Commentaires Positifs	Commentaires Négatifs
Commentaires Positifs Prédits	7 027	3 743
Commentaires Négatifs Prédits	2 973	6 257

TAB. 5 – *Matrice de confusion obtenue à partir du lexique General Inquirer.*

Afin de palier à ce problème, sans avoir recours aux techniques de Traitement Automatique des Langues qui pourraient être coûteuses au niveau de l'adaptation à ce type de corpus, nous avons tenté de tenir compte des négations de manière plus simpliste. Nous avons créé un troisième lexique contenant les négations et minimiseurs. Nous en avons répertorié six : *not, ever, never, less, no, badly*.

La détection, dans un commentaire, d'un terme appartenant à ce nouveau lexique inverse la polarité du prochain terme détecté appartenant aux deux autres lexiques (termes négatifs et termes positifs). Par exemple, la détection du mot "*not*" suivie du mot "*good*" soustraira 1 au score du commentaire alors qu'auparavant le score aurait été incrémenté. Nous pouvons voir

dans la matrice de confusion (tableau 6) que les résultats de la classification des commentaires négatifs sont légèrement améliorés par rapport à ceux de la première expérimentation. Par contre ceux concernant la classification des commentaires positifs sont dégradés. Les résultats globaux (voir troisième colonne du tableau 3) sont également moins bons que ceux obtenus avec la classification qui ne tient pas compte de la négation. Précisons que, cette fois-ci, les commentaires classés représentent 80% du corpus de test, soit 6% de plus que la première expérimentation.

	Commentaires Positifs	Commentaires Négatifs
Commentaires Positifs Prédits	7 057	3 132
Commentaires Négatifs Prédits	2 943	6 868

TAB. 6 – *Matrice de confusion obtenue à partir de nos lexiques en prenant en compte les négations.*

3.2.4 Conclusion de l'approche linguistique

La prise en compte des négations comme elle a été faite ici n'est pas réellement efficace et les meilleurs résultats obtenus correspondent à ceux de la première expérimentation, avec l'utilisation de notre dictionnaire, en tenant compte uniquement des adjectifs et verbes. Les résultats pourraient peut-être être améliorés en considérant les négations comme des mots négatifs Benamara et al. (2007). Une analyse relationnelle des phrases (qui permet d'extraire les relations existant entre les mots) serait certainement la solution la plus efficace, mais il faudrait pour cela un outil adapté au style de langage utilisé dans le corpus étudié ici.

Nous pouvons également remarquer que les résultats obtenus avec le dictionnaire construit à partir du corpus traité sont significativement supérieurs à ceux obtenus à partir du dictionnaire *General Inquirer*. On peut donc en déduire que si une application vraiment ciblée est envisagée (par exemple une classification de commentaires de films), l'utilisation d'un dictionnaire d'opinion adapté est plus recommandée que l'utilisation d'un dictionnaire général.

3.3 L'approche statistique

Nous nous intéressons ici à deux techniques d'apprentissage supervisé : les Machines à Vecteur Support (SVM) et les Classifieurs Naïfs Bayesiens (NB), dont une méthode de classification Naïve Bayesienne avec sélection de variables (SNB). Nous allons tout d'abord présenter rapidement les trois méthodes utilisées. Puis nous expliquerons comment nous avons procédé pour le choix des classes à prédire. Enfin, nous présenterons les résultats obtenus avec les trois techniques.

3.3.1 Classification Naïve Bayesienne

Le classifieur Bayesien Naïf a démontré son efficacité sur de nombreuses applications réelles (Hand et Yu, 2001). Cette méthode de classification repose sur l'estimation de la pro-

babilité d'occurence d'évènements avec la règle de Bayes. Ce classifieur, qui suppose que les variables explicatives sont conditionnellement indépendantes, nécessite uniquement l'estimation des probabilités conditionnelles univariées. Les études menées dans la littérature (Liu et al., 2002) ont démontré l'intérêt des méthodes de discrétisation pour l'évaluation de ces probabilités conditionnelles.

3.3.2 Classification Naïve Bayesienne avec sélection de variables : KHIOPS

Khiops [2] est un outil de data mining permettant de construire automatiquement un modèle de classification performant, sur de très grandes volumétries (Boullé, 2005, 2006, 2007). Dans une première phase de préparation de données, les variables explicatives sont évaluées individuellement au moyen d'une méthode de discrétisation optimale dans le cas numérique et de groupement de valeurs optimal dans le cas catégoriel. Dans la phase de modélisation, un modèle de classification est construit, en moyennant efficacement un grand nombre de modèles basés sur des sélections de variables. L'outil Khiops a été évalué avec succès lors de plusieurs challenges internationaux de data mining, et est utilisé à France Télécom par une trentaine d'utilisateurs pour des problèmes de marketing (churns, scores d'appétence...), de text mining, web mining, étude technico-économique, ergonomie, sociologie.

3.3.3 Machine à Vecteur Support : SVM^{light}

L'objectif des Machine à Vecteur Support est de trouver un hyperplan qui sépare les exemples positifs des exemples négatifs. SVM^{light} [3] est une implémentation de la Machine à Vecteur Support de Vapnik (Vapnik, 1995). Les algorithmes d'optimisation utilisés sont décrits dans Joachims (2002) et Joachims (1999a). Ce code a déjà été utilisé pour un grand nombre de problèmes, notamment la classification de texte (Joachims, 1999b, 1998), ainsi que des tâches de reconnaissance d'image, de bioinformatique et des applications médicales.

3.3.4 Choix des classes de commentaires

Comme nous l'avons précisé plus tôt, les commentaires composant le corpus d'analyse sont à l'origine notés sur dix classes (de 0,5 à 5). Il aurait été possible, au moins pour les méthodes NB et SNB, de réaliser la classification sur plus de deux classes. Mais l'augmentation du nombre de classes de projection rend l'interprétation des résultats d'autant plus difficile. En effet, imaginons une classification sur 5 classes de notes allant de 1 à 5, 1 désignant un commentaire très négatif et 5 un commentaire très positif. Le fait qu'un commentaire noté 1 par son auteur soit noté 2 par l'outil d'apprentissage doit-il être considéré comme une erreur, ou doit-on fixer une tolérance, et si oui, à combien doit-on la fixer. Pour éviter ces problèmes d'interprétation et faciliter l'analyse des résultats, nous avons donc décidé de réduire le nombre de classes.

Pour tous les tests qui vont suivre, nous avons utilisé la méthode SNB.

2. Outil téléchargable en shareware sur http ://perso.rd.francetelecom.fr/boulle/
3. Outil téléchargable en freeware sur http ://svmlight.joachims.org/

Nous avons tout d'abord tenté une projection sur deux classes : NEG pour négatif et POS pour positif. Nous sommes partis de l'hypothèse que les commentaires notés 0,5 et 1 étaient tous négatifs, et que les commentaires notés 5 étaient tous positifs. Nous avons donc réalisé la phase d'apprentissage uniquement sur les commentaires notés 0,5, 1 et 5, puis réalisé la classification indépendamment sur chaque classe restante. Les résultats sont présentés dans le tableau 7.

	1,5	2	2,5	3	3,5	4	4,5
NEG	78,18	75,46	65,58	53,62	33,58	28,2	17,4
POS	21,82	24,54	34,42	46,38	66,42	71,8	82,6

TAB. 7 – *Résultats de la projection sur deux classes.*

Nous pouvons observer que les résultats sont assez cohérents. Comme on pouvait s'y attendre, les plus mauvais résultats se situent au niveau des commentaires notés 2,5, 3 et 3,5. Nous avons donc tenté d'introduire une troisième classe (NEUTRE) en considérant, pour l'apprentissage, que les commentaires notés 3 appartenaient tous à cette nouvelle classe. La phase d'apprentissage se fait donc avec une classe négative contenant les reviews notées 0,5 et 1, une classe neutre contenant les reviews notées 3 et une classe positive contenant les reviews notées 5. Les résultats, qui sont présentés dans le tableau 8, montrent que, pour pratiquement toutes les classes, plus de la moitié des commentaires tombent dans la classe NEUTRE.

	1,5	2	2,5	3	3,5	4	4,5
NEG	40,6	32,76	20,88	14,92	9,64	9,72	8,28
NEUTRE	51,56	59,08	67,02	65,98	55,36	44,06	30,14
POS	7,84	8,16	12,1	19,1	35	46,22	61,58

TAB. 8 – *Résultats de la projection sur trois classes.*

Au vu de ces derniers résultats, nous avons donc décidé de partager le corpus en 2 classes. La classe positive regroupant les commentaires ayant une note comprise entre 3 et 5, et les commentaires négatifs contenant les commentaires notés de 0,5 à 2,5.

3.3.5 Résultats

Ici, nous avons décidé de n'avoir aucun *a priori* sur les données. Nous avons conservé les commentaires tels qu'ils ont été écrits par leurs auteurs avec seulement des prétraitements minimaux. Rappelons que les seuls traitements subis par le corpus sont la minusculisation des caractères, la suppression de la ponctuation ainsi que la suppression des mots n'apparaissant qu'une seule fois dans le corpus d'apprentissage. Nous n'avons appliqué sur le texte aucun traitement linguistique. Chaque commentaire est représenté sur un vecteur composé de 12 153

variables. Comme on peut le voir dans le tableau 9 qui présente les résultats des trois expérimentations, les meilleurs scores sont obtenus avec SVM^{light}, suivis par KHIOPS et enfin le classifieur naïf bayesien.

	KHIOPS	Naïve Bayes	SVM^{light}
Accuracy	76,3	69,8	79,6
Précision	76,6	71	81,6
Rappel	76,2	69,8	76,5
F_{score}	76,4	70,4	79

TAB. 9 – *Résultats des expérimentations sans prétraitement.*

3.3.6 Analyse des résultats obtenus avec la méthode SNB

Voici une analyse plus appronfondie des résultats obtenus avec l'outil KHIOPS, appliqué sur le corpus représenté en sac de mots, lors de la phase d'apprentissage, sans autres prétraitements que la minusculisation des caractères et la suppression de la ponctuation.

Le nombre de variables (mots différents) présentes dans le corpus d'apprentissage s'élève à 12 153. L'outil en a sélectionné 305 qui lui paraissent les plus informatives pour la classification d'opinion. L'étude de ces variables sélectionnées permet d'apprendre des informations sur le corpus.

La première constatation que l'on peut faire est que les 305 variables séléctionnées possèdent des degrés d'information très variables, et peu d'entre elles sont très informatives (voir figure 4). Les variables sont classés en fonction de leur *level*. Le level est directement relié à la probabilité *a posteriori* d'un modèle de discrétisation, avec une normalisation O-1. Il vaut 0 lorsque la variable n'est pas du tout informative, et 1 lorsque la variable a un niveau d'information maximal.

Cette liste de variables informatives, contient une majorité de mots que l'on peut catégoriser comme mots porteurs d'opinion (voir tableau 10), mais pas seulement. D'autres mots, dont la présence dans la liste est plus surprenante, sont aussi informatifs que les mots d'opinion, voir d'avantage. On peut par exemple voir dans le tableau 11 l'importance du mot "*and*". Le tableau se lit de la façon suivante. la première colonne donne la fréquence du mot "*and*" dans un commentaire, la deuxième colonne donne le nombre de commentaires contenant autant de fois le mot "*and*" qu'indiqué dans la première colonne. Enfin, les deux dernières colonnes nous présente comment sont réparties ces commentaires dans les classes positive et négative.

On peut donc observer que plus le mot "*and*" apparait dans un commentaire, plus la probabilité que ce commentaire soit positif est élevée. On peut donc penser que certains auteurs ont tendance à être plus prolixe lorsqu'ils ont apprécié un film que lorsqu'ils ne l'ont pas aimé. Cette impression se confirme avec la présence dans les variables informatives de mots usuels

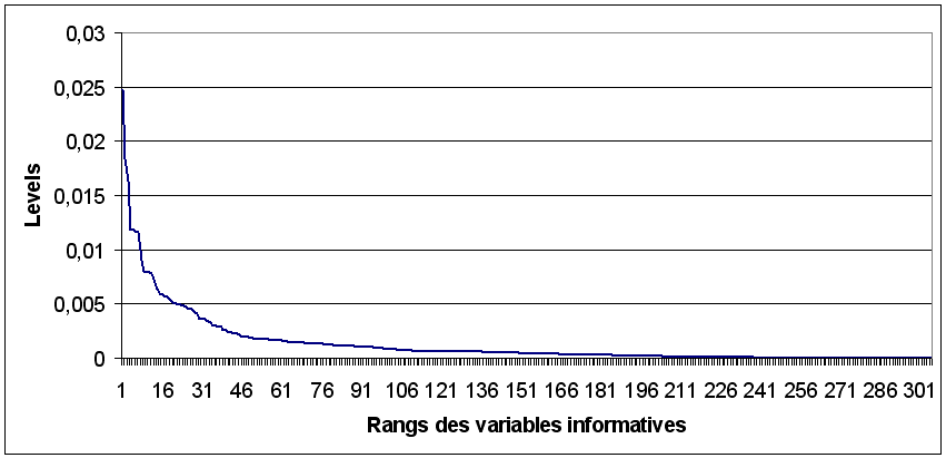

FIG. 4 – *Évolution du niveau d'information des variables séléctionnées par KHIOPS.*

Variables informatives	love, great, best, loved, boring, ok, brilliant, awesome, worst good, stupid, funny, amazing, hilarious, crap, awsome, hate bad, excellent, fantastic, favorite, ...

TAB. 10 – *Exemples de mots contenus dans la liste des variables informatives.*

tels que "*movie*" (tableau 12) ou "*film*" (tableau 13), ou encore le mot de liaison "*the*" (tableau 14).

D'autres types de mots sont aussi présents dans cette liste. Notamment les mots qui sont utilisés pour décrire le genre du film comme "*action*" (tableau 15) ou "*thriller*" (tableau 16). Les statistiques de ces mots démontrent une nouvelle fois que certains auteurs décrivent plus les films qu'ils ont apprécié que les autres.

Une autre information importante extraite de la liste de variables informatives est la présence de négations comme "*not*" (tableau 17) et "*didn't*" (tableau 18). Comme on pouvait s'y attendre, ces mots sont plus présents dans les commentaires négatifs que dans les commentaires positifs. Ceci explique pourquoi, dans l'approche linguistique, la détection des négations est une phase nécessaire à la classification des commentaires négatifs.

3.3.7 Conclusion sur l'approche statistique

Les meilleurs scores ont été obtenus avec la méthode SVM. Malgré tout, la méthode SNB reste très intéressante à utiliser car elle permet d'obtenir des informations sur les variables les

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	33 725	52,54%	47,46%
n = 1	4 439	41,17%	59,83%
1 < n < 5	1 619	29,96%	70,04%
n > 4	217	5,99%	94,01%

TAB. 11 – *Statistiques concernant le mot "and".*

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	30 849	53,54%	46,46%
n = 1	9 151	38,05%	61,95%

TAB. 12 – *Statistiques concernant le mot "movie".*

plus informatives, autrement dit celles qui permettent de classer les commentaires, que l'on n'aurait pas devinées instinctivement (importance du mot "*and*" par exemple).

3.4 L'approche hybride

Nous avons expérimenté une approche hybrique, mélant statistique et linguistique, en appliquant des prétraitements linguistiques sur le corpus afin de réduire le nombre de variables le représentant, et vérifier l'incidence qu'ils ont sur la classification. Nous avons également effectué une dernière expérimentation avec une représentation sous forme de tri-grammes de lettres, expérimentation qui n'appartient pas vraiment à l'approche hybride mais qui est une autre méthode de représentation des textes utilisée en Traitement Automatique des Langues.

3.4.1 Avec correction orthographique

Nous avons appliqué une correction orthographique (Guimier de Neef et al., 2002) sur tous les commentaires du corpus. L'espace de représentation est réduit à 10 783 variables, soit une réduction de plus de 11% par rapport à la représentation précédente. Nous pouvons voir dans le tableau de résultats (tableau 19) que cette réduction ne déteriore pas les résultats de la classification, certains d'entre eux sont même améliorés. Le F_{score} obtenu avec SVM^{light} gagne 1,3 point avec cette représentation par rapport à la représentation sans prétraitement.

3.4.2 Avec correction orthographique et lemmatisation

En plus de la correction orthographique, nous avons appliqué un deuxième traitement au texte, la lemmatisation (Guimier de Neef et al., 2002), qui consiste à remplacer tous les termes du texte par leur forme canonique (infinitif pour les verbes, et masculin singulier pour les autres

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	37 013	51,33%	48,67%
n = 1	2 987	33,58%	66,42%

TAB. 13 – *Statistiques concernant le mot "film".*

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	29 179	52,29%	47,71%
n = 1	8 923	45,77%	54,23%
n = 2	1 898	34,62%	65,38%

TAB. 14 – *Statistiques concernant le mot "the".*

mots). Le nombre de variables est de 9 095, soit 25% d'éléments en moins comparé à la représentation sans prétraitement. Les résultats présentés dans le tableau 20 montrent que cette nouvelle représentation obtient sensiblement de meilleurs scores, excepté pour le classifieur naïf bayesien.

3.4.3 Tri-grammes de lettres

Pour cette dernière expérimentation, nous avons voulu testé les trigrams de lettres imaginés par Shannon (1951). Les tri-grammes sont des sous-séquences de trois lettres extraites d'une séquence plus longue. Voici un exemple de représentation sous forme de tri-grammes :

Phrase de départ : " the movie "
Représentation en tri-grammes : " the|he |e m| mo|mov|ovi|vie "

Cette représentation ne correspond pas à un prétraitement linguistique mais à un mode de représentation différent des commentaires. Aucun hypothèse n'est faite sur le vocabulaire.

Le nombre de variables avec cette représentation s'élève à 9 722. Les résultats présentés dans le tableau 21 sont moins bons que ceux obtenus avec la représentation lemmatisée, et ce, pour tous les outils testés.

3.4.4 Conclusion sur l'approche hybride

Ces nouveaux résultats confirment les résultats obtenus avec l'approche uniquement statistique, à savoir que les scores obtenus avec la méthode SVM sont supérieurs aux autres, et ce, quelle que soit la représentation utilisée. Les meilleurs scores sont obtenus après correction et lemmatisation des textes. Les résultats obtenus avec la représentation en tri-grammes de lettres sont étonnament faibles car cette représentation est très performante en règle générale. Ceci est

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	39 262	50,51%	49,49%
n = 1	738	23,04%	76,96%

TAB. 15 – *Statistiques concernant le mot "action".*

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	39 725	50,29%	49,71%
n = 1	275	8%	92%

TAB. 16 – *Statistiques concernant le mot "thriller".*

sans doute lié au vocabulaire utilisé dans ce corpus qui est très pauvre. Pour preuve, le nombre de tri-grammes est inférieur au nombre de mots, alors qu'en général leur nombre est largement supérieur.

4 Conclusion

Nous avons testé et évalué différentes approches de classification d'opinion. La première approche consiste à construire un dictionnaire d'opinion manuellement avec l'aide de techniques simples de Traitement Automatique des Langues. Ce dictionnaire permet ensuite de classer les textes selon leur polarité, positive ou négative. Pour la seconde approche, ce sont des outils d'apprentissage automatique qui ont été utilisés dans le but, toujours, de classer les textes selon qu'ils expriment une opinion générale positive ou négative. Plusieurs outils d'apprentissage supervisé ont été comparés : un classifieur Naïf Bayesien, un classifieur Naïf Bayesien avec sélection de variables et une Machine à Vecteur Support. Pour finir nous avons appliqué des tâches linguistiques sur le corpus afin de changer la représentation des textes et comparer les résultats des classification effectuées avec les méthodes d'apprentissage.

Les expérimentations effectuées sur des commentaires de films issus de blogs d'opinion ont montré que les méthodes statistiques étaient plus performantes que notre approche linguistique. Cette dernière pourrait sans doute être améliorée avec une meilleure détection des négations. Cette approche reste cependant intéressante car elle nécessite peu d'exemples (textes déjà classés) par rapport aux méthodes statistiques pour lesquelles une quantité importante d'exemples est nécessaire pour mener à bien la phase d'apprentissage. Le classifieur le plus performant sur notre problème est le SVM, et ce, quelle que soit la représentation du texte. On retrouve ici les résultats de Joachims (1999b) ; le SVM étant très bien adapté pour aux données décrites en grande dimension et aux données creuses.

Le classifieur Naïf Bayesien avec sélection de variables est moins performant que le SVM en classification mais apporte des informations intéressantes sur l'aspect informatif du vocabu-

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	36 189	48,51%	51,49%
n = 1	3 811	64,13%	35,87%

TAB. 17 – *Statistiques concernant le mot "not".*

Fréquence du Mot	Commentaires Concernés	Commentaires Négatifs	Commentaires Positifs
n = 0	38 896	49,42%	50,58%
n = 1	1 104	70,47%	29,53%

TAB. 18 – *Statistiques concernant le mot "didn't".*

	KHIOPS	Naïve Bayes	SVM^{light}
Accuracy	76,6	69,7	80,1
Précision	76,8	71,1	83
Rappel	76,7	69,7	77,8
F_{score}	76,7	70,4	80,3

TAB. 19 – *Résultats des expérimentations avec la correction orthographique comme prétraitement.*

	KHIOPS	Naïve Bayes	SVM^{light}
Accuracy	76,4	69,1	81,1
Précision	77,3	70,4	82,9
Rappel	76,4	69,1	78,4
F_{score}	76,8	69,7	80,6

TAB. 20 – *Résultats des expérimentations avec la correction orthographique et la lemmatisation comme prétraitements.*

	KHIOPS	Naïve Bayes	SVM^{light}
Accuracy	76,2	58	79,5
Précision	76,5	69	83,7
Rappel	76,2	58	73,2
F_{score}	76,3	63	78,1

TAB. 21 – *Résultats des expérimentations avec la réprésentation sous forme de tri-grams de lettres.*

laire. Il fait émerger des mots pertinents pour l'opinion qu'aucune des autres approches n'aurait pu faire émerger, notamment des mots que l'on a *a priori* tendance à nettoyer comme "*and*" et "*the*".

Références

Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, et V. Subrahmanian (2007). Sentiment analysis : Adjectives and adverbs are better than adjectives alone. In *International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado, U.S.A, 26/03/2007-28/03/2007*, http ://www.aaai.org/Press/press.php, pp. 203–206. AAAI Press.

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2006). MODL: a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. (2007). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research 8*, 1659–1685.

Carbonell, J. (1979). *Subjective Understanding: Computer Models of Belief Systems*. Phd thesis, Yale.

Cohen, W. (1996). Learning trees and rules with set-valued features. In *In Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 709–716. AAAI Press.

Das, S. et M. Chen (2001). Yahoo! for amazon: Extracting market sentiment from stock message boards. *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.

Ding, X. et B. Liu (2007). The utility of linguistic rules in opinion mining. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 811–812. ACM.

Dziczkowski, G. et K. Wegrzyn-Wolska (2008). An autonomous system designed for automatic detection and rating of film reviews. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on 1*, 847–850.

Guimier de Neef, E., M. Boualem, C. Chardenon, P. Filoche, et J. Vinesse (2002). Natural language processing software tools and linguistic data developped by france télécom rd.

Hand, D. et K. Yu (2001). Idiot bayes ? not so stupid after all? *International Statistical Review 69*(3), 385–399.

Hatzivassiloglou, V. et K. R. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, pp. 174–181. Association for Computational Linguistics.

Hu, M. et B. Liu (2004a). Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 168–177. ACM.

Hu, M. et B. Liu (2004b). Mining opinion features in customer reviews. In D. L. Mcguinness, G. Ferguson, D. L. Mcguinness, et G. Ferguson (Eds.), *AAAI*, pp. 755–760. AAAI Press / The MIT Press.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, *Springer*.

Joachims, T. (1999a). Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, 169–184.

Joachims, T. (1999b). Transductive inference for text classification using support vector machines. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, San Francisco, CA, USA, pp. 200–209. Morgan Kaufmann Publishers Inc.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers.

Kanayama, H. et T. Nasukawa (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. EMNLP.

Kelly, E. et P. Stone (1975). *Computer Recognition of English Word Senses*.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, Morristown, NJ, USA, pp. 768–774. Association for Computational Linguistics.

Liu, B., M. Hu, et J. Cheng (2005). Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, New York, NY, USA, pp. 342–351. ACM.

Liu, H., F. Hussain, C. Tan, et M. Dash (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery 4*(6), 393–423.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, et K. J. Miller (1990). Introduction to wordnet: an on-line lexical database*. *Int J Lexicography 3*(4), 235–244.

Morinaga, S., K. Yamanishi, K. Tateishi, et T. Fukushima (2002). Mining product reputations on the web. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 341–349. ACM.

Nigam, K. et M. Hurst (2006). Towards a robust metric of polarity. In *Computing Attitude and Affect in Text: Theory and Applications*, pp. 265–279.

Pang, B. et L. Lee (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2*(1-2), 1–135.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Morristown, NJ, USA, pp. 79–86. Association for Computational Linguistics.

Pereira, F., N. Tishby, et L. Lee (1993). Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 183–190. Association for Computational Linguistics.

Riloff, E. et J. Wiebe (2003). Learning extraction patterns for subjective expressions. In

Proceedings of the 2003 conference on Empirical methods in natural language processing, Morristown, NJ, USA, pp. 105–112. Association for Computational Linguistics.

Schapire, R. E. et Y. Singer (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning 39*(2/3), 135–168.

Shannon (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, 50–64.

Stone, P. J., D. C. Dunphy, M. S. Smith, et D. M. Ogilvie (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

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

Turney, P. D. et M. L. Littman (2003). Measuring praise and criticism: Inference of semantic orientation from association.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Wilks, Y. et J. Bien (1984). Beliefs, points of view and multiple environments. In *Proc. of the international NATO symposium on Artificial and human intelligence*, New York, NY, USA, pp. 147–171. Elsevier North-Holland, Inc.

Wilson, T., J. Wiebe, et R. Hwa (2004). Just how mad are you? finding strong and weak opinion clauses. In *In Proceedings of AAAI*, pp. 761–769.

Yu, H. et V. Hatzivassiloglou (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Morristown, NJ, USA, pp. 129–136. Association for Computational Linguistics.

Summary

Community sites are by nature dedicated places to express and publish opinions. *www.flixster.com* is an example of participative web site, with dozens of millions of enthusiasts sharing their feelings/views on movies. A thorough study of this information richness would allow a better knowledge of Internet users, their expectations, their needs. In this purpose, a necessary step is the automatic opinion classification. In this paper we present three approaches allowing to classify texts according to the opinion they convey. The first approach is a linguistic approach who consists in labelling words carrying opinion thanks to NLP (Natural Language Processing) methods. These words allow to classify texts. The second approach is based on machine learning techniques. The third approach combines linguistic techniques, in order to clean the corpus, and statistic techniques, to classify each text.