



Université
de Toulouse



CNRS - INPT - UPS - UT1 - UTM

Master 2 Recherche en Informatique – Spécialité RIBD :

« Recherche d'Information et Base de données »

Responsable de la formation :

Professeur Mohand BOUGHANEM

Equipe Systèmes d'Information Généralisés / Filtrage et
Recherche d'Information

EXPÉRIMENTATION DE FONCTIONS POUR LA DÉTECTION D'OPINIONS DANS LES BLOGS

Mémoire soutenu

le 23 Juin 2010

par

Faiza Belbachir

Composition du jury

Président : Mme Nathalie Aussenac-Gilles

Rapporteur : Mme Farah Benamara

Directeur de recherche : Mohand Boughanem

Encadrant : Guillaume Cabanac, Missen Malik Muhammad Saad

Expérimentation de fonctions pour la détection d'opinions dans les blogs

MÉMOIRE

pour l'obtention du

Master Recherche de l'Université Paul Sabatier
(Spécialité Informatique)

par

Faiza Belbachir

Composition du jury

Président : Mme Nathalie Aussenac-Gilles
Rapporteur : Mme Farah Benamara
Directeur de recherche : Mohand Boughanem
Encadrant : Guillaume Cabanac, Missen Malik Muhammad Saad

Remerciements

Je tiens à remercier **M. Claude Chrisment**, responsable de l'équipe SIG, de m'avoir accueillie dans son équipe.

Je remercie aussi, **M. Zoubir Mammeri**, responsable du master Informatique et de Télécommunication, et **M. Mohand Boughanem**, responsable du parcours Recherche d'Information et Base de données, qui m'ont permis d'accéder au master.

Je remercie vivement **M. Guillaume Cabanac** de m'avoir encadrée et de m'avoir donné l'occasion de travailler sur ce riche et passionnant sujet. Sa responsabilité, son sens des relations, son expérience, sa rigueur et ses critiques constructives m'ont été très utiles pour avancer dans mon master.

Je remercie **M. Mohamed Misen Saad Malik** pour m'avoir aidée et soutenue durant tout mon master. Je tiens à remercier particulièrement mes parents et ma sœur sans lesquels ma vie n'aurait plus aucun sens.

*Je dédie ce mémoire
à mes proches...*

Table des matières

Introduction

1

Contexte de travail : la détection d'opinions dans les blogs

3

1.1	Introduction	3
1.2	Fouille d'opinions et analyse de sentiment	4
1.2.1	Définition	4
1.2.2	Applications	4
1.2.3	Difficultés de la fouille d'opinions et de l'analyse de sentiments	5
1.2.4	Processus de la fouille d'opinions	6
1.3	Les campagnes d'évaluation	7
1.3.1	TREC blog track	7
1.3.2	NTCIR	9
1.3.3	Conclusion	10

2

État de l'art de la détection d'opinions

11

2.1	Introduction	11
2.2	Détection de l'opinion	11
2.2.1	Approches basées sur le lexique (Lexicon-Based Approach)	11
2.2.2	Approches basées sur l'apprentissage machine (Machine Learning)	13
2.3	Classification de la polarité des opinions	16
2.3.1	Approches basées sur le lexique	16
2.3.2	Approches basées sur l'apprentissage machine (Machine Learning)	16
2.4	Bilan	18

3

Contribution : apport de la classification des requêtes pour la détection d'opinions et de leurs polarité **19**

3.1	Introduction	19
3.2	Détection d'opinion	20
3.2.1	Features utilisées	20
3.2.2	Combinaison des features	24
3.2.3	Détection de la polarité	28
3.2.4	Détection de polarité sans classification des topics	31
3.2.5	Détection de la polarité avec classification des topics	32
3.2.6	Bilan	36

Conclusion et perspectives **37**

Bibliographie **39**

Table des figures

1.1	Graphe indiquant l'importance de la détection d'opinions. [Esu06]	4
1.2	Processus de fouille d'opinions	6
1.3	Blog track 2006, opinion retrieval task, topic 871	8
2.1	Classement des universités qui travaillent sur la polarité	17
3.1	Features complémentaires	23
3.2	Impact de la pertinence dans la détection d'opinions	24
3.3	Tests de validation croisée	25
3.4	Combinaison des features	25
3.5	Résultats des différents classifieurs	25
3.6	L'évaluation des topics de 2007 en utilisant SVM	26
3.7	Topics qui donnent de bons résultats	27
3.8	Topics qui donnent les résultats les plus mauvais	27
3.9	Jugement des annotateurs pour les différents thèmes	29
3.10	Classification en thèmes des topics de blog 2006	30
3.11	Classification en thèmes des topics de blog 2007	30
3.12	Les features de la polarité	31
3.13	Résultats des différentes expérimentations de la polarité	32
3.14	Les groupes de la classification des topics	33
3.15	Comparaison des résultats de la polarité	33
3.16	Polarité positive	34
3.17	Polarité négative	34
3.18	Détection d'opinions avec classification	35

Introduction

De nos jours plusieurs applications et plateformes sur le web nous permettent de déposer des avis, de partager des sentiments et des opinions sur une variété de sujets. Vue l'importance de ces informations dans plusieurs domaines (politique, commercial ou individuel), il serait important de déterminer l'information subjective contenue dans les textes. Mais la détection automatique d'opinions et l'analyse des sentiments sont confrontées à des problèmes qui la distinguent de la recherche thématique traditionnelle, car le sentiment est exprimé de manière très variée et très subtile. De ce fait, plusieurs travaux de recherche se sont intéressés à ce problème. En règle générale, il existe deux types d'approches pour la détection d'opinions et l'analyse du sentiment, celles qui sont basées sur le lexique et celles qui sont basées sur l'apprentissage machine.

Les approches basées sur le lexique utilisent un dictionnaire de mots subjectifs. Ce dictionnaire peut être général comme par exemple « General Inquiry, Sentiwordnet, Opinion Finder, NTU », ou il est déduit à partir du corpus étudié. Plusieurs techniques sont utilisées pour détecter la subjectivité des mots. On peut citer celle de l'analyse par la conjonction entre les adjectifs (qui prédit l'orientation de la subjectivité des adjectifs en analysant les couples d'adjectifs liés par « and », « or », « but », « either-or » ou « neither-or »), celle de l'analyse par les relations lexicales ou celle de l'analyse par glossaire.

Les approches basées sur l'apprentissage machine consistent à attribuer des données à un classifieur pour l'apprentissage. Ce dernier génère un modèle qui est utilisé pour la partie test de l'apprentissage. Ce type d'approche comprend deux aspects : extraction de features et apprentissage du classifieur. Les principales features utilisées sont : mots seuls, bigrammes, tri-grammes, part of speech et polarité. Les principaux classifieurs sont les SVM, Naive Bayes, Maximum Entropy et régression logistique. La plupart des chercheurs considèrent que seuls les adjectifs sont porteurs de subjectivité, alors que d'autres pensent que certains adverbes, noms et verbes peuvent aussi contenir de la subjectivité.

Afin d'évaluer les résultats des chercheurs dans le domaine, plusieurs campagnes d'évaluation ont vu le jour telles que TREC, NTCIR etc. Ces campagnes proposent une collection de test qui comporte un ensemble de documents, un ensemble de topics (sujets) et des jugements de pertinence « relevance judgement ». Ces derniers représentent la perception humaine. Les résultats sont donc comparés au « qrels » à travers des mesures tels que le rappel, la précision, la MAP, etc. Nous nous intéressons dans notre travail au problème de la détection d'opinions et de la polarité dans les blogs. Le blog est un site web constitué de billets (blogposts) écrits par une personne selon un rythme périodique. Il est généralement créé par le blogueur lui-même à travers une plateforme ou d'un logiciel de publication. Pour évaluer nos résultats, nous utilisons la collection test « blog06 » proposé par TREC, et le logiciel TREC_EVAL. Pour la détection d'opinions, nous proposons une approche mixte basée sur le lexique SentiWordNet et sur l'apprentissage

machine. Nous utilisons un certain nombre de features dont les principales sont l'émotivité, la subjectivité, la réflexivité et l'adressage. Nous combinons ces features de différentes façons afin de trouver la combinaison qui permet de détecter au mieux l'opinion dans les blogs.

Nous nous intéressons ensuite à la détection de la polarité en déterminant si l'opinion est positive, négative ou neutre. Nous utilisons la même approche que celle proposée pour la détection d'opinions. Les features utilisées sont : le nombre de mots positifs, le nombre de mots négatifs, le nombre de mots neutres et le nombre d'adjectifs se trouvant dans le document. Nous proposons une catégorisation des topics et étudions son impact sur la détection de la polarité.

Le document est divisé en trois chapitres. Le premier chapitre introduit le problème de la détection d'opinions et l'analyse des sentiments en mentionnant les domaines d'applications et les difficultés associées. Les campagnes d'évaluation TREC et NTCIR sont présentées dans ce chapitre. Le deuxième chapitre présente quelques approches dans le domaine, en particulier celles qui ont participé aux campagnes d'évaluation TREC ou NTCIR. Le troisième chapitre expose notre contribution. L'approche proposée est présentée ainsi que l'évaluation des résultats qui a été faite sur la collection de « blog06 » proposée TREC. Enfin le quatrième chapitre conclut ce mémoire et présente les perspectives à nos travaux.

Chapitre 1

Contexte de travail : la détection d'opinions dans les blogs

1.1 Introduction

Le Web a beaucoup évolué au cours de ces dernières années sur plusieurs aspects, et principalement avec l'intégration des multimédia comme les photos et les vidéo. Il permet plus d'échanges qui n'étaient pas possibles auparavant. À sa création le Web ressemblait à un grand livre composé de pages reliées entre elles par des liens hypertextes. L'internaute était restreint au seul rôle de « consommateur » ne pouvant que consulter les informations. Toutefois, avec l'évolution des technologies (haut débit, infrastructure) et surtout avec l'arrivée du web social, l'internaute passe d'un statut de consommateur à un statut d'acteur, et on parle ainsi de web social.

Le web social fait référence à une vision d'Internet considérée comme un espace de socialisation, un lieu où les personnes peuvent interagir entre elles, et non pas uniquement échanger des documents. A partir de 2005, le web social a connu un développement très important avec l'arrivée du web 2.0. Plusieurs travaux ont été faits dans ce domaine comme par exemple [CCC⁺08]. Le web social est associé à différents systèmes sociaux tel que le réseautage, les blogs, et autres. Le réseautage social est l'ensemble des moyens mis en œuvre pour relier des personnes. On peut citer par exemple les réseaux sociaux professionnels, les services en ligne d'anciens (anciens élèves de grandes écoles ou université, etc.), les sites de rencontres.

Le blog est un site web constitué de billets (blogposts) écrits par une personne selon un rythme périodique (tous les jours, chaque semaine, etc.). Le blog est généralement créé par le blogueur lui-même à travers une plateforme ou d'un logiciel de publication. Souvent des utilisateurs font des commentaires sur des blogs. Un blog peut référencer d'autres blogs qui constituent sa blogoliste. L'univers des blogs est appelé blogosphère et est considéré comme un réseau social le plus populaire [DF04]. La présence d'informations de nature subjective apparaît de manière très visible dans les blogs. Les bloggeurs postent des commentaires, font part de leurs sentiments, et diffusent leurs opinions sur divers sujets. Ces opinions ont une grande importance dans plusieurs applications et on assiste ainsi à un nouveau concept de « fouille d'opinions et analyse de sentiments ». Nous présentons dans ce qui suit quelques applications de fouille d'opinions. Nous montrons ensuite les difficultés de l'extraction automatique d'opinions sur le web. Nous décrivons le processus de fouille d'opinions (acquisition du corpus, analyse du corpus et évaluation des résultats). Nous terminons ce chapitre par la présentation des campagnes d'évaluation TREC et NTCIR.

1.2 Fouille d'opinions et analyse de sentiment

1.2.1 Définition

L'information textuelle dans le monde peut être classée en deux catégories principales, les faits et les opinions. Les faits sont un énoncé objectif sur les entités et les événements dans le monde. Les opinions sont subjectives et reflètent les sentiments des gens ou des perceptions au sujet des entités et des événements. La plupart des recherches existantes se sont portées sur la fouille et l'extraction de faits, par exemple, la recherche d'information, la recherche sur le Web et beaucoup d'autres. On assiste, ces dernières années, à une prise de conscience de l'importance de l'opinion sur le web, ce qui explique les nombreux et récents travaux dans ce domaine [HT09]

La figure 1.1 montre que la fouille d'opinions est une discipline récente et a été sujette à un grand intérêt ces dernières années [Esu06].

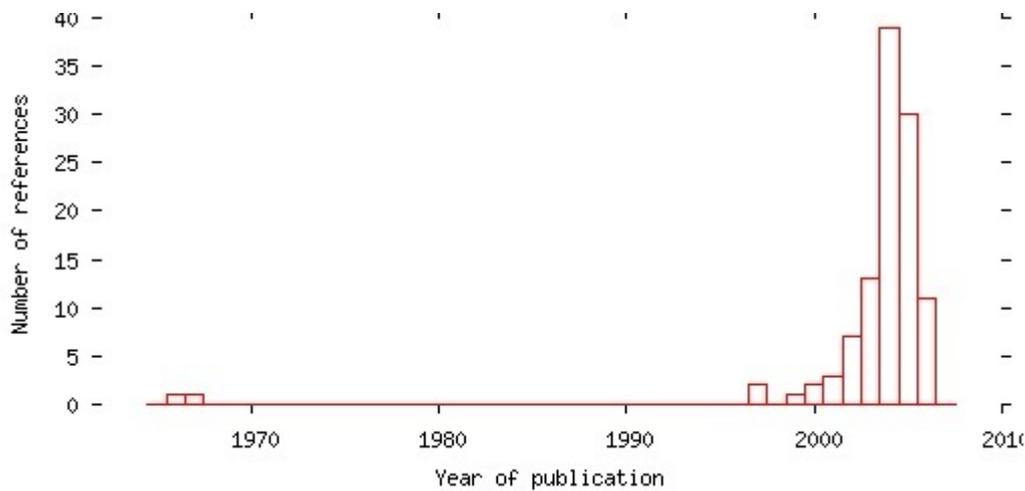


FIGURE 1.1 – Graphe indiquant l'importance de la détection d'opinions. [Esu06]

1.2.2 Applications

L'importance de la détection d'opinion est présente dans plusieurs domaines et plusieurs applications ont vu le jour dans ce contexte [Liu07]. Nous citons brièvement quelques applications ci-dessous.

– *la politique*

avant de promulguer une nouvelle loi, les politiciens essaient de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur tel homme politique pour une élection présidentielle.

– *les entreprises*

à travers la fouille d'opinions, les entreprises peuvent connaître l'opinion des internautes sur leurs produits et savoir par exemple pourquoi leur produit ne se vend pas malgré le prix bas et la qualité. Ceci peut être dû à l'emballage ou la couleur ou autre. Cette connaissance leur permettra d'améliorer leurs produits.

– *les clients*

les sondages montrent que la majorité des internautes font des recherches d'avis sur un produit et sont prêts à payer plus cher un produit dont l'avis est plus favorable qu'un autre.

1.2.3 Difficultés de la fouille d'opinions et de l'analyse de sentiments

L'opinion et le sentiment sont le plus souvent décrits par la polarité. Cette dernière est en générale, soit positive (opinion favorable), soit négative (opinion défavorable), soit neutre. Nous montrons ci-dessous quelques difficultés de cette fouille d'opinions [PL08] :

- difficulté due à l'ambiguïté des mots. Par exemple le mot « petit » est un fait dans la phrase suivante « il est petit ». Par contre il exprime une opinion dans « c'est un petit ».
- difficulté due à la structuration de la phrase. Par exemple on oppose deux parties d'une phrase avec la conjonction « mais », par exemple l'histoire du film est intéressante mais les acteurs étaient mauvais. Dans ce cas la polarité de la deuxième partie est opposée à la première.
- difficulté due au contexte : par exemple dans la phrase « je trouve que le film est excellent mais ma sœur le trouve mauvais », les deux opinions sont données par des personnes différentes. Dans l'exemple suivant « J'ai regardé le film Avatar dans la salle de cinéma 21 qui est très jolie », l'opinion est donnée sur la salle et non sur le film. Dans le dernier exemple « les acteurs du film ont bien joué, la musique est bonne mais je n'ai pas apprécié ce film », l'opinion de la dernière partie de la phrase est la plus importante.
- difficulté due au vocabulaire qu'on utilise pour exprimer une opinion. Il diffère d'une personne à une autre, comme par exemple un anglo-saxon lorsqu'il exprime ses sentiments utilise des mots bien représentatifs de ce qu'il ressent contrairement aux personnes qui ne connaissent pas ou peu sa langue.
- difficulté due à l'emploi d'une thématique. Une même thématique peut être utilisée dans différentes classes et peut exprimer une toute autre signification, comme par exemple « un vieux vin », ou un « vieil habit ».
- difficulté due au langage qu'utilisent les internautes pour s'exprimer. Les ponctuations ne sont pas forcément utilisées pour marquer les fins de phrases, des mots spécifiques sont utilisés tel que : « ha ha ha », « Good », « super ».
- difficulté de déterminer un lexique adapté à l'analyse de l'ensemble des textes d'opinion.
- difficulté à trouver une association entre l'opinion et la requête. En effet une opinion dans un document ne porte pas forcément sur la requête considérée.

La section suivante décrit le processus de fouille d'opinions.

1.2.4 Processus de la fouille d'opinions

Le processus d'un système de fouille d'opinions comprend trois étapes : acquisition et analyse du corpus, étude de la pertinence des documents par rapport à un sujet, détection de l'opinion et ré-ordonnement des documents. La figure 1.2 nous montre les étapes de la fouille d'opinions

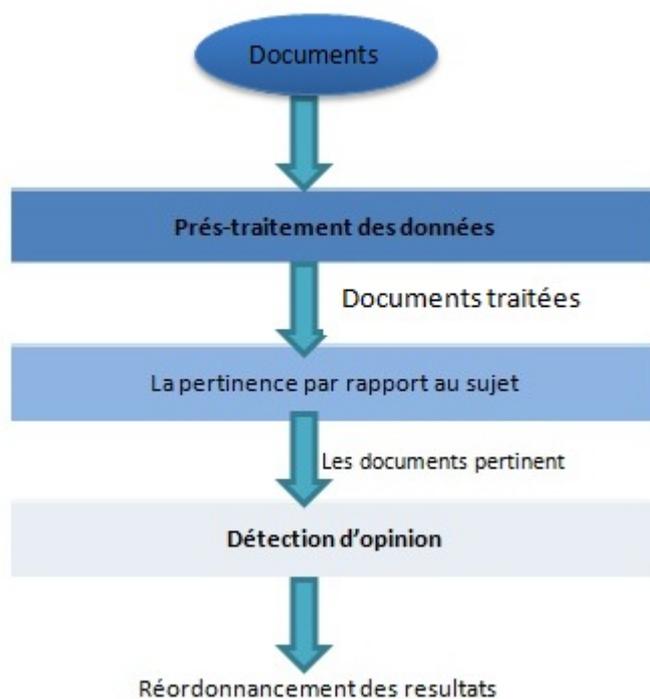


FIGURE 1.2 – Processus de fouille d'opinions

1. *Acquisition et prétraitement du corpus*

Dans cette phase, les textes sont prétraités linguistiquement. Une élimination des mots vides et des mots qui n'apportent aucune information est faite, ainsi qu'une analyse lexicale pour enlever les mots qui ont un sens commun (redundant). Concernant les blogs la majorité des travaux se sont portés sur un nettoyage au niveau des balises HTML inutiles, ainsi que sur les documents qui ne sont pas écrits en langue anglaise. Certaines techniques de fouille d'opinions [MB09] utilisent la structure des phrases afin d'identifier l'opinion. Dans cette étape, un étiquetage grammatical est fait (pour reconnaître l'adjectif, l'adverbe, le verbe, etc.), les grammaires de dépendances sont utilisées pour structurer la phrase de manière hiérarchique.

2. *La pertinence par rapport au sujet*

Cette étape consiste à étudier la pertinence des documents par rapport à un sujet donné, appelé « topic » dans Trec. Nous employons indifféremment ces deux termes. Parmi les méthodes les plus utilisées, on peut citer la méthode probabiliste OKAPI et BM25 [JWR00]. Les documents sont classés, et généralement les 1000 premiers documents les plus pertinents sont extraits, et sont utilisés pour l'étape suivante : la détection d'opinions.

3. La détection d'opinions

Plusieurs méthodes ont été utilisées pour la détection d'opinions. Leur but est de réordonner les documents pertinents selon un score d'opinion. Ainsi les documents qui contiennent le plus d'opinions sont classés parmi les premiers.

Après ces étapes, une évaluation des résultats est faite. Les résultats sont souvent confrontés à la perception humaine de l'opinion. La comparaison est faite grâce à des mesures de similarité. Plusieurs campagnes d'évaluation ont vu le jour, permettant aux chercheurs de présenter leurs travaux et de les évaluer sur des collections test élaborées par ces campagnes. Nous présentons ci-dessus les campagnes TREC et NTCIR.

1.3 Les campagnes d'évaluation

1.3.1 TREC blog track

1.3.1.1 Définition

Le mot TREC signifie « Text Retrieval Conference » et désigne l'ensemble des conférences organisées par le NIST « National Institute of Standard and Technology » sur la recherche d'information. Plusieurs tâches ont fait l'objet de recherches dans ces conférences, dont le Blog Track qui a été introduit en 2006. Chaque année, de nouvelles tâches sont définies, telles que :

1. la recherche de billets de blogs pertinents (baseline adhoc blog post retrieval task) ;
2. la recherche d'opinion dans les blogs (opinion finding blog post retrieval task) dont les travaux ont été présentés dans TREC 2006 ;
3. la détection de la polarité qui consiste à déterminer si l'opinion est positive, négative ou neutre sur un sujet (polarised opinion finding blog post retrieval task). Cette tâche a été introduite pour la première fois dans TREC 2007 [MOS07] ;
4. La recherche de blog dont le principal intérêt porte sur un sujet (blog finding distillation task). Cette tâche a été introduite pour la première fois dans TREC 2007 et traitée aussi dans TREC 2008 [OMS08] ;
5. la recherche de blog dont le principal intérêt porte sur un sujet, et tenant compte de certaines facettes (Faceted blog distillation task). Trois facettes ont été spécifiées pour TREC 2009 [MG10]. La première porte sur l'opinion (la valeur de cette facette est égale à « opinionated » ou bien à « factual »). La deuxième facette porte sur le caractère personnel ou officiel des documents recherchés (la valeur de la facette est « personal » ou « official »). La troisième facette est égale à « in depth » si l'analyse sur le topic est importante autrement elle est égale à « shallow ». Cette tâche a été aussi proposée dans TREC2010 ;
6. L'identification des premières actualités pertinentes à une date donnée (Top stories identification task) et sélection des blogposts pertinents pour ces actualités. Cette tâche a été introduite dans TREC 2009. Une catégorisation des nouvelles a été proposée pour TREC 2010 : international, national, politique, sport, technologie, business, science).

Plusieurs universités et laboratoires de recherche participent à ces conférences TREC et utilisent les collections de test proposées par TREC pour évaluer leurs résultats.

1.3.1.2 Collection de test

Une collection de test comporte trois parties [MC08] un ensemble de documents, un ensemble de topics et les jugements de pertinence « relevance judgements ». Le résultat d'une exécution d'un système de recherche exécutant une tâche sur une collection de test est appelé un « run ».

1. Document

L'ensemble des documents d'une collection de test doit refléter la réalité : diversité des sujets, des formats et des styles. Dans le blog track, deux collections ont été créés par l'université de Glasgow. La première dite « blog06 » collection et a été utilisée dans TREC2006, 2007 et 2008. C'est un échantillon des blogs de la blogosphère. Les documents ont été collectés pendant 11 semaines du 6 décembre 2005 au 21 février 2006. Cette collection est de 148 Go et se compose de 38,6 Go de feed (blogs), de 88,8 Go de permalink (un billet de blog simple avec tous les commentaires associés) et 20,8 Go de home page (page d'accueil d'un blog). La deuxième collection dite « blog08 » [MG10] a été créée en 2008 et a été utilisée dans TREC 2009. Elle est plus volumineuse que la précédente, et a été collectée de janvier 2008 à février 2009. La taille de cette collection est de 2309 Go et contient 808 feeds, 1445 permalink et 56 homes pages. Chaque document est caractérisé par un unique identifiant appelé DOCNO.

2. Topic

Les topics sont généralement construits spécialement pour la tâche pour laquelle ils sont utilisés. Un topic peut être une entité nommée comme un nom d'une personne, un emplacement ou un organisme mais il peut être un concept comme un type de technologie. Un topic est caractérisé par un identifiant (num), un titre, une description (généralement une phrase décrivant le topic) et une narration qui décrit l'information que l'individu recherche et qu'il considère comme pertinente. La campagne d'évaluation propose chaque année de nouveaux topics. 50 topics ont été définis en 2006, 50 autres en 2007 et 50 topics en 2008. La figure 1.3 illustre un exemple de topic de TREC 2006.

```

<top>
<num> Number: 901 </num>

<title> jstor </title>

<desc> Description:
Find opinions on JSTOR, the system developed to make scholarly
journals available from a digital archive.
</desc>

<narr> Narrative:
Reports of difficulty or ease in using JSTOR are relevant opinions. A
statement that one is lucky to have access or wishes to have access to
JSTOR is a relevant opinion. A statement that information is available
in JSTOR is not an opinion. Simply citing JSTOR as a reference for a
document is not an opinion.
</narr>

</top>

```

FIGURE 1.3 – Blog track 2006, opinion retrieval task, topic 871

3. Jugements de pertinence

Les jugements de pertinence ne peuvent pas être faits manuellement pour deux raisons majeures. La première est que la collection de documents est trop grande et cela prendrait énormément de temps pour juger toute la collection. La deuxième raison est que le jugement est subjectif. Il diffère d'un juge à un autre et aussi pour un même juge à des moments différents. Pour remédier à cela, TREC utilise la technique du pooling. Cette dernière consiste à choisir un certain nombre de runs soumis par les participants, ensuite à choisir parmi ces runs les n (généralement $n = 100$) premiers documents en éliminant les doublons. Cet ensemble de documents (appelé «pool») est évalué par des assessseurs humains et le résultat nommé « qrels » contient la vérité terrain.

1.3.1.3 Évaluation de l'efficacité des systèmes de fouille d'opinions dans TREC

Comme mentionné précédemment, la campagne d'évaluation TREC propose chaque année plusieurs tâches. Pour qu'un participant puisse prendre part à une tâche, il doit fournir aux organisateurs de l'évaluation au moins un run. Ce dernier représente une liste de documents restitués pour chaque topic traité et classés par pertinence décroissante (généralement les 1000 premiers documents). Le run soumis sera évalué par rapport aux «qrels». Cette évaluation est faite par le logiciel TREC_EVAL¹, développé par Chris Buckley à travers des mesures d'évaluation. Les mesures les plus utilisées sont le rappel (qui est le ratio entre le nombre de documents pertinents retrouvés et le nombre total de documents pertinents), la précision (qui est le ratio entre le nombre de documents pertinents retrouvés et le nombre total de documents retrouvés), la MAP (Mean Average Precision qui est la moyenne arithmétique des précisions moyennes calculée sur l'ensemble des topics de la tâche considérée), la R-Precision (R-Prec), Binary Preference (Bpref) et les mesures de haute précision (P@5, P@10, P@15). Pour plus d'informations sur ces mesures, le lecteur peut se reporter à [CHBC10].

1.3.1.4 Participants

Plusieurs universités et laboratoires de recherches participent à ces conférences TREC pour les tâches proposées (un groupe peut participer à plusieurs tâches). Dix sept groupes ont participé à TREC2006, vingt-quatre à TREC2007 (vingt pour la fouille d'opinions, onze pour la polarité et neuf pour le blog distillation), vingt groupes dans TREC2008 (vingt pour la baseline, dix-neuf pour la fouille d'opinions, seize pour la polarité et douze pour le blog distillation), et à TREC2009, neuf groupes ont participé à «Faceted blog distillation» et sept groupes à «Top stories identification».

1.3.2 NTCIR

L'Institut National en Informatique (NII) du Japon organise chaque année le workshop NTCIR (NII Test Collection for Information Retrieval System) dans le domaine de la recherche d'information. La tâche de l'analyse d'opinions a été l'objet d'étude de NTCIR-6 [YS07] et NTCIR-7 [Lia]. Elle est proposée aussi pour NTCIR-8 qui aura lieu en juin 2010. La détection d'opinions se fait au niveau des phrases. Quatre sous-tâches ont été définies dans NTCIR-6 :

- savoir si une phrase contient une opinion ou pas (opinionated sentence judgement) ;
- extraire le nom de la personne ou de l'entité qui émet l'opinion (opinion holder extraction) ;
- savoir si une phrase est pertinente ou non pour un topic (relevance sentence judgement) ;

1. http://trec.nist.gov/trec_eval/

- détecter la polarité.

Une cinquième sous-tâche a été ajoutée dans NTCIR-7 et qui consiste à extraire la cible sur laquelle porte l'opinion.

1.3.2.1 La collection de test

La collection de NTCIR-6 est multilingue (chinois traditionnel, japonais, anglais). Elle est constituée de documents newswires, récoltés entre 1998 et 2001. Les documents en japonais ont été pris à partir de journaux japonais tels que Yomiuri et Nainichi, et sont au nombre de 490 et composés de 15279 phrases. Les documents en chinois ont été pris à partir de journaux chinois tels que China Times, United Daily News etc, et sont au nombre de 843 et composés de 11907 phrases. Tandis que les documents en anglais, ils ont été pris des journaux tels que Mainichi Daily News, Korea Times, etc. Ces documents sont au nombre de 439 et sont composés de 8528 phrases.

La collection de NTCIR-7 [Lia] a considéré une quatrième langue qui est le chinois simplifié, et la collection a été enrichie par d'autres documents. Un topic est constitué d'un numéro, d'un titre, d'une description, d'une narration et d'un champ précisant la pertinence. Dans NTCIR-6, trente-deux topics ont été définis pour le chinois, trente pour le japonais et vingt-huit pour l'anglais, tandis que dans NTCIR-7, vingt-deux topics ont été définis pour le japonais, dix sept pour l'anglais, dix-sept pour le chinois traditionnel et seize pour le chinois simplifié.

1.3.2.2 Jugements de pertinence

Les jugements de pertinence ont été faits manuellement. Pour le chinois, sept annotateurs ont été désignés et chaque document est jugé par 3 annotateurs. Trois annotateurs ont jugé tous les documents japonais et trois autres les documents en anglais. Dans NTCIR-7, le nombre d'annotateurs a augmenté.

1.3.2.3 Participants

Dans NTCIR-6, cinq groupes ont participé pour le chinois traditionnel, six groupes pour l'anglais et trois groupes pour le japonais. D'autres nouveaux groupes ont participé à NTCIR-7.

1.3.3 Conclusion

Vu le nombre de tâches proposées et le volume de la collection, TREC reste sans contexte la référence en matière d'évaluation des systèmes de recherche d'information. Ceci est confirmé par le nombre de participants aux campagnes d'évaluation de TREC. Nous présentons dans le chapitre suivant les travaux des principaux participants à TREC et quelques travaux présentés à NTCIR.

Chapitre 2

État de l'art de la détection d'opinions

2.1 Introduction

Nous nous intéressons dans ce chapitre aux travaux relatifs à la détection d'opinions et de la polarité (la classification des sentiments). La détection d'opinions est une tâche qui permet d'extraire les opinions d'un ensemble de documents pertinents pour un sujet donné. Elle est confrontée à des problèmes qui la distinguent de la recherche traditionnelle thématique dont les sujets sont souvent identifiés par des mots clés seulement. L'opinion (ou le sentiment) peut être exprimée de manière très variée et subtile et donc il est difficile de la déterminer. La classification du sentiment (polarité) est une sous-tâche de la détection d'opinions. Elle consiste de façon générale à déterminer si l'opinion du document sur le sujet est positive ou négative. La détection d'opinions se fait au niveau du document, du paragraphe ou de la phrase. Le travail présenté dans [MBC09a] montre les difficultés supplémentaires dans le cas de la phrase.

Nous présentons dans la première section de ce chapitre les travaux relatifs à la détection d'opinions et dans la deuxième section ceux relatifs à la détection de la polarité.

2.2 Détection de l'opinion

En général, il existe deux types d'approches utilisées, l'une basée sur un lexique et l'autre sur l'apprentissage machine.

2.2.1 Approches basées sur le lexique (Lexicon-Based Approach)

Ces approches utilisent un lexique de mots qui contiennent un sentiment. Ce lexique est soit externe c'est-à-dire construit indépendamment de tout corpus, il peut être général (SWN, SentiWordNet², SUBJ lexique, General Inquiry, Wilson lexicon) ou construit manuellement, soit généré automatiquement à partir du corpus (les mots qui contiennent une opinion sont extraits directement du corpus). À chaque mot du lexique est associé un ensemble de scores d'opinions. Ce score est traité différemment par les différentes approches pour le calcul du score d'opinion d'un document. La méthode la plus simple est de donner à un document un score d'opinion égal au nombre total de mots qui contiennent une opinion présents dans le document.

2. <http://sentiwordnet.isti.cnr.it/>

L'apprentissage peut être efficace pour générer des scores individuels pour les documents et aussi pour l'apprentissage d'une fonction de classement qui combine le score d'opinion et le score de pertinence dans une seule fonction de classement.

Nous exposons brièvement dans ce qui suit les travaux de quelques universités et laboratoires de recherche qui se sont basés sur le lexique pour déterminer l'opinion. Le résultat de ces approches est comparé à ceux jugés corrects (on nomme ces derniers « baseline »).

1. *L'université de Glasgow*

Le travail de cette université a été exposé dans l'article [HMP⁺07] et présenté à TREC 2007. Deux approches pour la recherche d'opinion ont été proposées. La première est basée sur un lexique de termes pondérés et la deuxième utilise Opinion Finder [Opi05].

Dans la première approche, un dictionnaire de 12000 mots anglais dont les termes sont pondérés a été utilisé. Ce dictionnaire est construit à partir de la collection de documents. Le poids des termes est calculé en utilisant un modèle de pondération des termes. En plus, la proximité des termes de la requête par rapport aux phrases avec opinion est prise en considération pour améliorer la performance dans la recherche d'opinion. À cet effet, le modèle pBiL2 « terme proximité » est utilisé. Ce dernier est un modèle binomial aléatoire qui calcule le score d'une paire de termes de la requête dans le document. Dans la deuxième approche, les auteurs utilisent OpinionFinder. Le score d'opinion du document est calculé en fonction de celui fourni par OpinionFinder. Les résultats montrent que la première approche améliore les performances dans la recherche d'opinion de 15,8 % de celles du « baseline » et de 8,96 % de celles obtenues par OpinionFinder.

2. *L'université d'Indiana*

Les travaux de cette université sur la fouille d'opinions sont exposés dans l'article [YYZ07] et présentés à TREC 2007. Les auteurs utilisent cinq lexiques. Le premier HF « High Frequency » regroupe les termes avec opinion, les plus couramment utilisés. Ce lexique est construit semi-automatiquement. Le deuxième lexique WL « Wilson's Lexicon » est construit à partir des termes de Wilson's subjectivity. Il est divisé en trois collections : subjectivité forte, subjectivité faible et les termes qui amplifient l'intensité de l'opinion. Le troisième lexique LF « Low Frequency » est construit à partir de mots créés par les gens pour exprimer leur forte opinion (exemple « soooo », « goood ») et sont des mots rares. Le quatrième lexique est appelé « IU » lexique (I and You). Les auteurs remarquent que les pronoms (I, you, my, your, our, me) sont souvent utilisés pour exprimer des opinions. Ce lexique est constitué par les n -grammes qui commencent et qui se terminent par un terme IU. Le dernier lexique dit « OA » lexique (Opinion Acronym) est construit manuellement. Le score d'un document est calculé selon une pondération des scores des différents lexiques. Cette approche améliore les performances du « baseline » de 14,07 %.

3. *Le laboratoire DUTIR (Information Retrieval laboratory of Dalian University of Technology)*

Ce laboratoire a participé à toute les tâches de TREC 2007 Blog Track [STS⁺07]. Dans la recherche d'opinion, les auteurs créent leur propre lexique composé de 2000 mots subjectifs qui sont fréquents dans le corpus étudié, au lieu d'utiliser les lexiques tels que General Inquiry ou SentiWordNet. Ces derniers, d'après eux, contiennent certains mots qui ne sont pas fréquemment utilisés dans les blogs. La requête est étendue par des mots pris du descriptif ou de la narration du topic, ou bien de Wikipedia. L'expression d'opinion dans le blog est détectée par une simple recherche de mots subjectifs du lexique considérée autour des mots de la requête. L'évaluation de cette approche a donné une amélioration de 10,38 % par rapport à la « baseline ».

4. *L'université d'Amsterdam*

Cette université a participé à la campagne de TREC 2007 dans plusieurs tâches notamment celle de la recherche d'opinion [Mdr06]. Les auteurs utilisent dans ce contexte-là une approche basée sur le lexique « General Inquiry » qui contient 10000 mots et est construit manuellement. Il est divisé en 5 catégories. Catégorie « positifs-négatifs », catégorie « émotionnelle » (pleasure, pain, feel, etc.), catégorie « pronom » (self, our, you), catégorie « adjectif » (adjectif relationnel, adjectif indépendant) et catégorie « respect » (une liste complète des catégories de General Inquiry est donnée dans <http://www.wjh.harvard.edu/~inquirer/homecat.htm>). Pour chaque blog post, deux valeurs « sentimentales » sont calculées, en utilisant les mots du dictionnaire de chaque catégorie : la valeur sentimentale au niveau du post « post opinion level » et la valeur sentimentale au niveau du blog « feed opinion level ». Dans les deux cas, la valeur d'opinion est le nombre d'occurrences des mots de chaque catégorie, normalisé par le nombre total de mots. La différence est le texte utilisé pour compter le nombre d'occurrences. Pour le « post opinion level », les phrases pertinentes sont extraites à partir du post, et utilisées pour le calcul de la valeur d'opinion. Les phrases pertinentes sont toutes les phrases qui contiennent le topic et celles qui les entourent immédiatement. Pour la deuxième valeur « feed opinion level », le feed auquel appartient le blogpost est utilisé. Les auteurs pensent qu'un feed contenant une grande quantité d'opinion contient vraisemblablement une opinion dans n'importe quel post. Les auteurs tiennent compte aussi de la qualité du blogpost, en particulier en filtrant les spams et en calculant un degré d'autorité qui correspond au nombre de liens relatifs à ce blog. Le score final d'opinion d'un document est une combinaison linéaire des scores pondérés, obtenus par les différentes méthodes : pertinence, opinion (blogpost et feed blog), autorité.

5. *L'université de Pohang (république de la Corée) : KLE (Knowledge and Language Engineering)*

Pour la détection d'opinions, Les auteurs proposent une approche basée sur le lexique. Ce dernier est créé en utilisant SentiWordNet et « Amazon's review corpus ». L'approche est basée sur le modèle « passage level » qui a été aussi utilisé pour la tâche de recherche de documents pertinents. Cette approche consiste à segmenter le document en passages, et à calculer le score d'opinion de chaque passage. Le passage ayant le score maximal est sélectionné. Les auteurs considèrent ensuite la partie du document qui contient ce passage et l mots avant et après ce passage dans le document. Le score est calculé pour cette partie et représente le score du document. Le modèle OKAPI est utilisé pour la normalisation des blogposts selon leur longueur. Le score d'opinion est calculé en utilisant une extension de la requête initiale par un ensemble représentatif de tous les mots subjectifs. Cet ensemble est appelé POW (Pseudo Opinionated Word). Cette université a participé à TREC 2008, et le travail a été présenté dans [LNK⁺08].

2.2.2 Approches basées sur l'apprentissage machine (Machine Learning)

Ces approches utilisent des classifieurs. Des données sont fournies au classifieur pour l'apprentissage. Les données représentent des phrases subjectives (ou des documents avec opinion). Le classifieur génère un modèle, qui sera utilisé dans la partie test. Des « features » sont utilisées pour l'apprentissage tels que les bigrammes, les n -grammes, POS (étiquettes morphosyntaxiques) etc. Plusieurs types de classifieurs ont été utilisés [PLV02] : SVM, Naive Bayes, Multiples Classifieur, Naïfs de Bayes, ainsi que la régression logistique.

1. Naive Bayes est une approche probabiliste qui utilise une loi de Bayes où les probabilités

sont fonction des mots contenus dans les documents :

$$P(c/d) = P(c) * \frac{P(d/c)}{P(d)} \quad (2.1)$$

Avec d est un document et c la classe du document. « $P(c/d)$ » est déterminé par le classifieur Bayésien naïf [YH03].

2. L'approche SVM repose sur la notion d'hyperplan séparateur et de marge maximale. Un hyperplan séparateur entre deux ensembles de points (ensemble de documents de polarité positive et l'ensemble de document de polarité négative) est la frontière entre ces deux ensembles. La marge représente la distance entre un de ces ensembles et cet hyperplan.
3. La régression logistique est une méthode statique permettant de produire un modèle pour décrire des relation entre une variable catégorielle et un ensemble de variables de prédiction.

2.2.2.1 Travaux utilisant l'apprentissage machine

1. Université d'Arkansas à Little Rock

Cette université a participé à la tâche de détection d'opinions et de polarité dans les blogs [ZJB07]. Trois approches différentes ont été proposées. La première est basée sur l'hypothèse que la présence d'un indicateur de subjectivité près du topic ou des mots de la requête permet de dire que l'opinion du document est sur le topic donné. Les mots indicateurs de subjectivité considérés sont : «I, You, We, Me, They, He, She» nommés IU model, et les mots «Like, Feel, Think, Love, hate, Suck, Nice, Good, Bad, awesome, awful, never, think, feel» nommé IU2. La recherche de ces indicateurs de subjectivité se fait sur une fenêtre de 20 mots. Deux runs ont été soumis, l'un utilisant seulement le titre du topic, l'autre le TDN (titre, narration, descriptif) et l'expansion de la requête avec des mots indicateurs d'opinion. La deuxième approche utilisée est basée sur l'apprentissage machine, de type SVM. Dans cette approche les topics sont divisés en 6 catégories : thing, compagny, food, events, location, person. Un SVM est utilisé pour chaque catégorie de topic. La troisième approche est basée sur le traitement du langage naturel (NLP) et dite «one-pass-processing approach» et elle traite la pertinence et la détection d'opinion en une seule étape. Les documents sont analysés et segmentés en passages. La première méthode qui est basée sur les IU mots donne de meilleurs résultats, elle améliore la «baseline» de 14,07 %, le CML de 2,96%, le IU2 de 3,2% , le TDN de 6,8% et le NPL de 5,3 %.

2. L'université de Neuchâtel (Suisse) (Computer Science Department)

Cette université a participé aux campagnes d'évaluation de TREC et NTCIR à travers deux travaux différents :

- (a) Les travaux de cette université présentés à TREC 2008 sont exposés dans l'article [FS08]. L'opinion et la polarité sont traitées en une seule étape. Le document est classé en positif, négatif, mixte ou neutre. Si le document est classé neutre alors il est considéré sans opinion autrement il est considéré comme contenant une opinion. Deux approches, basées sur la méthode de Muller sont utilisées pour cette classification. La première dite «Additive Model» utilise les statistiques d'un terme pour calculer un score de polarité pour chaque document. Ce score est basé sur le Z-score de Muller. La deuxième approche utilise le classifieur de la régression logistique, douze features et le Z-score de Muller. La requête est étendue de deux manières. La première est basée sur la méthode de Rocchio où les m termes les plus importants extraits des k premiers documents retrouvés pour la requête initiale, sont rajoutés à la requête étendue. La

deuxième extension utilise Wikipedia. Le titre du topic est soumis à Wikipedia et les dix premiers mots les plus fréquents du premier article retrouvé sont ajoutés à la requête.

- (b) Les travaux présentés à NTCIR-7 (MOAT : Multilingual Opinion Analysis Task) sont exposés dans l'article [OZ08]. Ils consistent à étudier l'opinion et la polarité au niveau d'une phrase dans un contexte multi-langues (anglais, japonais et chinois). L'approche utilisée est basée sur l'apprentissage machine. Les auteurs utilisent la méthode de Muller pour calculer les poids des termes, et le classifieur de type régression logistique pour déterminer la catégorie de la phrase (positive, négative, neutre ou sans opinion). Pour la langue anglaise, les auteurs étendent la requête initiale avec 500 mots identifiant les événements du discours (« explained », « commented », etc.) ou des expressions subjectives (« sympathized », « accused », etc.). Les auteurs remarquent que les résultats ne sont pas performants pour certaines requêtes.

3. *Le laboratoire « National Laboratory of Pattern Recognition », Chinese Academy of Sciences, Pékin de Chine*

Ce laboratoire a participé à NTCIR-7 pour la tâche MOAT (Multilingual Opinion Analysis Task), plus particulièrement aux sous-tâches de détection de phrases subjectives et d'extraction du porteur d'opinion (opinion holder extraction), pour la langue chinoise simplifiée. L'approche utilisée pour la détection d'opinions est basée sur l'apprentissage machine. Les auteurs utilisent un classifieur subjectif et un certain nombre de features tels que les adjectifs et verbes apparaissant dans la phrase, les entités nommées, la structure de la phrase lors de l'analyse, les idiomes apparaissant dans la phrase, et les mots de deux dictionnaires. Le premier dictionnaire « opinion operator lexicon » est constitué de verbes qui peuvent signaler un événement d'opinion (par exemple, « believe », « say », « persist », etc.) et est constitué à partir du corpus NTCIR-6. Le deuxième dictionnaire (opinion word lexicon) contient les mots (verbes, adjectifs et adverbes) qui expriment une opinion. Ce dictionnaire est construit à partir du corpus de NTCIR-6, du dictionnaire de sentiment de l'université nationale de Taiwan (il contient 11,088 mots) et du dictionnaire HowNet (qui contient 8,938 mots subjectif). Les auteurs concluent que les résultats ne sont pas satisfaisants et que d'autres features doivent être ajoutées.

Il existe d'autres travaux importants concernant le domaine de la détection d'opinions. Dans [ABM09], les auteurs proposent une approche pour l'analyse d'opinions dans les textes, basée sur une analyse lexicale-sémantique des expressions et sur les relations rhétoriques entre ces expressions. Les auteurs proposent une catégorisation des expressions d'opinions en quatre catégories « reporting expressions », « judgment expressions », « advise expressions », « sentiments expressions ». Chaque catégorie est divisée en sous-catégories. Le deuxième apport de ce travail est l'utilisation des relations rhétoriques entre les expressions contenant des opinions. Cinq types de relations rhétoriques ont été définies : « contrast », « correction », « explanation », « result », « continuation ». L'approche proposée consiste à segmenter le texte en segments contenant des opinions, et à représenter chaque segment par une structure contenant la catégorie à laquelle appartient le segment, la modalité associée, l'entité qui a émis l'opinion, le sujet de l'opinion, le mot d'opinion et le contenu de l'opinion. Cette représentation permet en utilisant un ensemble de règles de calculer l'opinion associée à un texte. La validation de cette approche s'est faite sur trois types de corpus : un corpus sur des critiques de film (pris de Telerama et AlloCine.fr), un corpus sur des lettres aux éditeurs (The San Francisco Chronicle et la Depeche du Midi) et un corpus sur des nouvelles (Le monde, 20 Minutes et MUC6).

En plus de la détection d'opinions (subjectif, objectif), des travaux on consisté à la classification de ces opinions par (positif, négatif, neutre).

2.3 Classification de la polarité des opinions

La classification des sentiments est un raffinement de la détection d'opinions dans la mesure où elle permet de classifieur les documents ayant une opinion sur un sujet en classes. Il existe deux types de classifications : binaire ou multi-classes. La classification binaire définit deux classes : positive et négative. Par contre la classification multi-classes définit cinq classes : fortement positive, positive, neutre, négative, fortement négative. La plupart des travaux se sont focalisés sur la classification binaire mais la classification multi-classes peut être utile dans certaines applications où on veut faire une meilleure classification. Les auteurs de [KS05] montrent qu'il est crucial d'utiliser des exemples neutres dans l'apprentissage de la polarité pour diverses raisons. L'apprentissage des exemples positifs et négatifs seuls ne permet pas de bien classifieur les exemples neutres. De plus, l'apprentissage d'exemples neutres facilite la distinction entre les exemples positifs et les exemples négatifs. Comme pour la détection d'opinions, il existe deux approches pour la polarité, l'une basée sur le lexique, l'autre sur l'apprentissage machine. Nous présentons dans ce qui suit quelques travaux pour cette tâche.

2.3.1 Approches basées sur le lexique

1. *L'université de Glasgow*

Le calcul de la polarité a été fait en utilisant leur première approche de la détection d'opinions. Cette approche est basée sur le même lexique interne de 12000 mots anglais. Une mesure de divergence est calculée des mots du dictionnaire qui sont distribués dans les documents pertinents. Le résultat a été exprimé par une mesure R_ACC qui est de 0,2295 et la valeur de la précision à 10 documents est de 0,37. Les résultats obtenus sont parmi les meilleurs [OMS08].

2. *L'université d'Indiana*

La même approche est utilisée pour la polarité que celle qui a été utilisée pour la détection d'opinions et les mêmes lexiques. Cette approche est basée sur l'utilisation du lexique WL « Wilson's Lexicon » et d'autres lexiques construits.

3. *L'université de Pohang (république de la Corée) : KLE (Knowledge and Language Engineering)*

La même approche pour la détection de l'opinion est utilisée pour la polarité sauf qu'au lieu de considérer la subjectivité du mot, ils considèrent l'orientation sémantique (polarité). Un score de polarité du document est calculé, et qui est la différence entre le score positif et le score négatif. Si cette différence dépasse un seuil, alors la polarité du document est considérée comme positive, si elle est inférieure au seuil, elle est considérée comme négative, autrement elle est neutre. Les auteurs considèrent seulement le titre du topic pour la requête.

2.3.2 Approches basées sur l'apprentissage machine (Machine Learning)

1. *L'institut « Institute of Computing Technology » de l'académie chinoise des sciences de Pékin de Chine*

Cet institut a participé à TREC 2007 dans la recherche d'opinion. Le travail est exposé dans

l'article [LCW⁺07]. Les auteurs utilisent le classifieur Drag-push pour obtenir la polarité des blogs pertinents. L'apprentissage de ce classifieur a été fait en utilisant les résultats de TREC 2006. Ils proposent aussi de filtrer les spams. Un blog est classifié comme un spam si le nombre de ses liens dépasse un seuil prédéfini. Ce seuil est calculé par une approche heuristique. Les SVM sont utilisés comme outil pour classifieur les blogs spam et les blogs non spam.

2. Le laboratoire DUTIR (Information Retrieval laboratory of Dalian University of Technology)

Concernant la polarité, les chercheurs utilisent une méthode mixte basée sur un lexique de 2000 mots subjectifs et sur un classifieur de type SVM. Ce dernier prend comme « features » les mots du lexique et aussi des mots sélectionnés par une méthode basée sur le gain de l'information. L'ensemble d'apprentissage n'est pas un ensemble de documents mais un ensemble de phrases contenant les mots du topic. Ces phrases sont extraites des documents. Les auteurs soulignent que les meilleurs résultats sont obtenus en utilisant seulement le titre du topic.

L'évaluation de la méthode de détection de polarité a donné comme valeur 0,3080. Nous présentons dans le tableau 2.1 quelques résultats de TREC 2007 concernant la détection de la polarité.

Group	Run	R-Acc	A@10	A@1000
UIUC (Zhang)	uic75cpnm	0.2295	0.3700	0.0493
UAmsterdam (de Rijke)	uams07ipolt	0.1827	0.2640	0.0418
IndianaU (Yang)	oqsnr2optP	0.1799	0.2800	0.0401
DalianU (Yang)	DUTRun2P	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	0.1510	0.2380	0.0427

FIGURE 2.1 – Classement des universités qui travaillent sur la polarité

Les universités qui ont participé a TREC 2007 sont classées par ordre décroissant de leurs mesure R_ACC.

2.4 Bilan

Nous avons présenté dans ce chapitre les principaux travaux sur la détection d'opinions et la polarité. Les approches utilisées sont basées sur un lexique ou sur l'apprentissage machine. Les travaux de la première approche se basent sur des lexiques externes (généraux tel que SentiWorNet ou construit manuellement, tel que le lexique LF), ou des lexiques internes construits à partir du corpus (tel que le lexique HF). Ces lexiques contiennent des mots subjectifs. Certains travaux ne s'intéressent qu'aux adjectifs, d'autres ajoutent des adverbes, des noms et des verbes.

La recherche de l'opinion du document sur un sujet consiste dans la plupart des travaux à chercher les mots subjectifs autour des mots de la requête.

Les travaux de la deuxième approche se différencient par le type du classifieur utilisé, et par le choix des features. Les features les plus couramment utilisées sont le mot seul, l'unigramme, le bigramme, les étiquettes morpho-syntaxique (POS). Les résultats de ces travaux montrent que pour que cette approche soit efficace, il faut que la base d'apprentissage soit conséquente et que le choix des features soit pertinent.

Afin d'améliorer les performances, certains travaux combinent l'utilisation du lexique avec l'apprentissage machine. Souvent, la requête initiale (formée par le titre du topic) est étendue par l'ajout de mots pris par exemple du descriptif et/ou de la narration et/ou de Wikipédia. La plupart des travaux ont montré que les résultats de la détection d'opinions ou de la polarité sont plus performants en utilisant seulement le titre du topic. Les résultats de TREC 2008 montrent que les travaux basés sur le lexique donnent de meilleurs résultats pour la détection de la polarité que ceux basés sur l'apprentissage machine et que la bonne performance d'un système de recherche d'opinion dépend de la bonne performance du système de recherche des documents pertinents utilisé. L'université d'Arkansas a montré que l'utilisation de la catégorisation des topics et les c-SVM (SVM par catégorie) donne de meilleurs résultats que sans catégorisation. Cela indique que la catégorisation des topics avant de les traiter est une piste de recherche à approfondir.

À partir de ces remarques, nous proposons dans le chapitre 3 une approche mixte, utilisant le classifieur de la régression logistique et un ensemble de features proposé dans l'article [MBC09b]. Nous proposons également une catégorisation des topics pour la détection d'opinions et de la polarité.

Chapitre 3

Contribution : apport de la classification des requêtes pour la détection d'opinions et de leurs polarité

3.1 Introduction

Dans le chapitre précédent, nous avons discuté des différentes approches liées à la détection d'opinions et de la polarité. En règle générale, toutes les approches sont soit basées sur le lexique, soit basées sur l'apprentissage machine.

Il existe des travaux qui utilisent la dépendance du topic dans leur recherche d'opinions (ils utilisent le titre du topic et/ou la description et/ou la narration) [MdR06] alors que d'autres proposent de ne pas tenir compte du topic, et on parle de recherche d'opinions indépendante du topic [VB07, VB08]. L'avantage des approches basées sur l'indépendance par rapport au topic est qu'elles sont générales et s'adaptent à n'importe quelle requête alors que les approches basées sur les topics ont la capacité de mieux comprendre le besoin d'information de l'utilisateur vu qu'elles visent directement ce qu'il recherche. Il a été observé que les approches qui se basent sur le topic donnent de meilleures performances que les approches qui sont indépendantes du topic.

Nous proposons, dans notre approche, un ensemble de features qui sont basées sur l'indépendance par rapport à la requête. Nous évaluons différentes combinaisons de ces features pour la détection d'opinions et de la polarité, et en utilisant trois classifieurs (régression logistique, SVM, Naive Bayes). Nous proposons aussi une classification des topics, et analysons son impact dans la détection des opinions et de la polarité.

Nos contributions portent sur :

- une proposition de features indépendants du topic qui nous permettent de détecter des opinions ;
- la détermination d'une combinaison de features pour mieux détecter les opinions ; ainsi une comparaison de l'efficacité des trois classifieurs sur la tâche de détection d'opinion en utilisant la meilleure combinaison a été faite ;
- une classification des topics de TREC blog des années 2006 et 2007 ;
- une analyse des effets de la classification sur la détection de polarité.

Dans la suite du chapitre, nous décrivons les détails de chacune de nos contributions.

3.2 Détection d'opinion

La tâche de détection d'opinion consiste à récupérer des documents qui sont pertinents pour une requête donnée et qui contiennent en plus des avis. Dans cette section, nous décrivons notre approche basée sur l'indépendance par rapport au topic pour la détection d'opinions et dont les particularités sont :

- elle ne nécessite pas de commentaires de l'utilisateur pour déterminer les documents pertinents ;
- elle n'utilise pas des corpus externes dans l'apprentissage des mots ou des phrases pour déterminer les opinions ;
- elle est basée sur le lexique et sur l'apprentissage machine.

Dans ce qui suit nous détaillons toutes les features d'indépendance de topic que nous avons utilisées.

3.2.1 Features utilisées

Nous proposons d'expérimenter 43 features présentés dans [MBC09b] pour la tâche de détection d'opinions dont les détails sont donnés ci-dessous. Notons que seules les features principales ont été abordées ici :

1. Émotivité-1

Les chercheurs ont exploité la présence des adverbes et des adjectifs dans un document comme un indicateur qui permet de déterminer les opinions. Nous calculons l'émotivité d'un document en comptant le nombre des adverbes et des adjectifs dans ce document.

$$Emot(d) = \frac{|\{w \in d | type(w) \in \{adj, adv\}\}|}{|\{w \in d | type(w) \in \{verb, noun\}\}|} \quad (3.1)$$

2. Émotivité-2

Cette feature est la somme du score Émotivité-1 sur le nombre total de mots du document qui est représenté par $|\{t \in d\}|$.

$$Emot2(d) = \frac{Emot(d)}{|\{t \in d\}|} \quad (3.2)$$

3. Subjectivité

Cette méthode utilise le lexique SentiWordNet. Dans cette ressource lexicale, chaque mot a plusieurs synonymes nommés « synsets ». Chaque « synset » est associé à trois scores numériques $Obj(s)$, $Pos(s)$ et $Neg(s)$, qui signifient objective, positive et négative des termes contenus dans le « synset ».

La méthode utilisée pour développer des SentiWordNet est détaillée dans l'article [sentiwordnet]. Chacun des trois scores varie de 0,0 à 1,0 et leur somme est de 1,0 pour chaque « synset ». Cela veut dire que la signification d'un « synset » est attribuée au score le plus élevé des trois. Par exemple le « synset » « Estimable (3) », correspondant au sens « peut être calculé ou estimé » de l'adjectif estimable, a un score Obj de 1,0 (et les scores Pos et Neg de 0,0), tandis que le synset « Estimable (1) » correspondant au sens « Dignes de

respect ou en haute estime » a un score positive de 0,75, un score négative de 0,0 et un score objectif de 0,25.

Dans notre méthode l'utilisation de cette ressource nous a permis d'estimer chaque mot d'un document selon l'algorithme suivant. Chercher chaque mot d'un document dans SentiWorNet. Áchaque mot t est associé plusieurs synsets (plusieurs sens s) donc plusieurs valeurs de positifs, de négatifs ainsi que de neutres. Puis, une heuristique est calculée, nommée subjective, par cette équation :

$$Subj(t) = \frac{\sum_{s \in t} neg(s) + pos(s)}{|t|} \quad t = \{s_1, \dots, s_n\} \quad (3.3)$$

Où $|t|$ est le nombre de sens de ce terme trouvé dans le lexique de SentiWordNet. La subjectivité d'un terme t dans un document est utilisée par quatre fonctions que nous avons proposées. Elles sont discutées ci-dessous :

– *Subjectivité-1*

$$Subj(d) = \frac{\sum_{t \in d} Subj(t)}{|\{t \cap t' | t \in d, t' \in SWN\}|} \quad (3.4)$$

$Subj(t)$ est la subjectivité d'un terme t tel qu'il est calculée en utilisant l'équation (3.3), et $|\{t \cap t' | t \in d, t' \in SWN\}|$ est le nombre total des termes trouvés dans le document qui appartiennent à SWN.

– *Subjectivité-2*

$$Subj(d) = \sum_{t \in d} \frac{\max \left(\sum_{t \in SWN}^{t \in d} pos(t), \sum_{t \in SWN}^{t \in d} neg(t) \right)}{|\{t \cap t' | t \in d, t' \in SWN\}|} \quad (3.5)$$

$|\{t \cap t' | t \in d, t' \in SWN\}|$ est le nombre total des termes trouvés dans le document(d) et qui appartiennent à SWN. Dans cette fonction, nous préférons ne donner un score positif à un document que si le nombre de termes positifs l'emportent dans le document et vice versa.

– *Subjectivité-3*

$$Subj(d) = \frac{\sum_{t \in d} Subj(t)}{|\{t \in d\}|} \quad (3.6)$$

C'est la subjectivité-1 normalisée en la divisant par le nombre total de mots dans le document d .

– Subjectivité-4

$$Subj(d) = \frac{\sum_{t \in d | Subj(t) \geq 0,5} Subj(t)}{|\{t \in d | Subj(t) \geq 0,5\}|} \quad (3.7)$$

La subjectivité-4 est identique à celle de subjectivité 3 sauf que $|\{t \in d | Subj(t) \geq 0,5\}|$ est le nombre total des termes du document ayant une valeur $\geq 0,5$ subjectivité.

Chacune des valeurs ci-dessus de la subjectivité est utilisée pour trois catégories de termes, (adjectifs, adverbes et verbes).

4. *Réflexivité*

Les blogueurs utilisent beaucoup de pronoms réflexifs comme « je l'ai, moi, moi-même » lors de l'écriture. Par exemple, l'utilisation de « I » dans « I think » « Je pense que », « my point of view is that », « mon point de vu est que », etc. Toutes ces phrases font référence à une opinion opinion, et par conséquent, nous incluons la mesure de la réflexivité. L'idée est que tout document avec un plus grand nombre de ces mots sera plus subjectif par rapport à celui qui a moins de nombre de ces mots. Cette mesure est exprimée par $Ref(d)$.

$$Ref(d) = \frac{|\{w \cap w' | w \in d, w' \in R\}|}{|R| + |A|} \quad (3.8)$$

$|\{w \cap w' | w \in d, w' \in R\}|$ est le nombre de pronoms réfléchis dans le document d qui appartiennent à la liste de réflexivité R que nous avons construit, et $|R|$ est le nombre total de pronoms dans la liste de réflexivité R , et $|A|$ est l'ensemble de nombre total de pronoms dans la liste d'adressage.

5. *Adressage*

La plupart des phrases trouvées dans les blogs contiennent les mots suivants «you, your, yours, yourself, yourselves, u, he, she, they, himself, herself, themselves» car les utilisateurs écrivent des commentaires sur un sujet, en s'adressant aux autres blogueurs. De ce fait l'utilisation de ces pronoms d'adressage est très fréquent. Par conséquent, nous considérons que la composante d'adressage dans le cadre de notre détection d'opinions, est comme suit :

$$Add(d) = \frac{|\{w \cap w' | w \in d, w' \in A\}|}{|A| + |R|} \quad (3.9)$$

$|\{w \cap w' | w \in d, w' \in A\}|$ représente le nombre d'occurrences des termes d'adressages dans le document d qui appartiennent à la liste d'adressage $|A|$ que nous avons préparé, et $|A|$ est égal au nombre total de pronoms dans la liste d'adressage A , et $|R|$ est le nombre total de pronoms dans la liste de réflexivité.

Les autres features que nous avons utilisées sont décrites dans le tableau 3.1.

Nom	Total-Nom	Nombre total des noms du document
	Avg-Nom	$\text{Tot-N}/(\text{tot-Ajd}+\text{tot-Adv}+\text{tot-V}+\text{tot-N})$
Adjectif	Total-Adjectifs	Nombre total d'adjectifs du document
	Avg-Adjectifs	$\text{Tot-Adj}/(\text{tot-Ajd}+\text{tot-Adv}+\text{tot-V}+\text{tot-N})$
Verbe	Total-verbes	Nombre total des verbes du document
	Avg-verbes	$\text{Tot-v}/(\text{tot-Ajd}+\text{tot-Adv}+\text{tot-V}+\text{tot-N})$
Adverbe	Total-adverbe	Nombre total d'adverbes du document
	Avg-Adverbes	$\text{Tot-Adv}/(\text{tot-Ajd}+\text{tot-Adv}+\text{tot-V}+\text{tot-N})$
Polarité	Avg-mots-possitive	$\sum \text{mot}(\text{pos})/\sum \text{mot}(\text{pos})+\sum \text{mot}(\text{neg})+\sum \text{mot}(\text{ntr})$
	Avg-mots-negative	$\sum \text{mot}(\text{neg})/\sum \text{mot}(\text{pos})+\sum \text{mot}(\text{neg})+\sum \text{mot}(\text{ntr})$
	Avg-mot-neutre	$\sum \text{mot}(\text{ntr})/\sum \text{mot}(\text{pos})+\sum \text{mot}(\text{neg})+\sum \text{mot}(\text{ntr})$

FIGURE 3.1 – Features complémentaires

Nous avons utilisé, en plus des features, un ensemble de phrases qui contiennent des opinions et qui ont été fréquemment utilisées dans les blogs de la collection de TREC 2006.

Nous prenons aussi en considération les scores de pertinence. Les scores de pertinence sont utilisés comme une caractéristique qui permettra d'avoir en premier rang les documents pertinents qui contiennent le plus d'opinion. Le tableau 3.2 compare les résultats obtenus en utilisant seulement les features par rapport aux résultats obtenus en utilisant la pertinence des documents et les features. Ces résultats nous montrent l'importance du score de pertinence dans la détection d'opinions.

		Avec Rank et Score		Sans Rank et Score	
Feature	Sub-Feature	MAP	P@10	MAP	P@10
Emotivité	Emotivité_1	0,3754	0,526	0.1965	0.2120
	Emotivité-2	0,3209	0,464	0.1438	0.1880
Sujektivité	Subjectivité-1	0,375	0,536	0.1515	0.1940
	Subjectivité-2	0,3728	0,542	0.1459	0.1700
	Subjectivité-3	0,3801(0,44%)	0,536	0.1540	0.1940
	Subjectivité-4	0,3879(2,51%)	0,546	0.1620	0.1760
Adressage	Avg_d'adressage	0,3787(0,07%)	0,538	0.1879	0.2080
	Nombre-tot-adressage	0,3787(0,08%)	0,538	0.2213	0.3320
refflexivité	Avg_refflexivité	0,3718	0,528	0.2051	0.2880
	Total-reflexive	0,3718	0,528	0.2051	0.2880
phrase-opi	avg-phrase-opinion	0,3567	0,496	0.1990	0.2120
	Total-phrase-opinion	0,3792(0,21%)	0,534	0.2054	0.3080
Adverbe	Avg-adverbe	0,3475	0,528	0.2091	0.2440
	Total-adverbe	0,3794	0,538	0.1895	0.2000
Nom	Avg-Nom	0,3858(1,95%)	0,55	0.2112	0.2560
	Total-Nom	0,3776	0,538	0.1895	0.200
Verbe	Avg-verbe	0,3885(2,66%)	0,578	0.2115	0.2440
	Total-verbe	0,3785(0,02%)	0,534	0.1845	0.1920
Adjective	Avg-Adjectives	0,354	0,494	0.1012	0.0830
	Total-Adjective	0,3764	0,534	0.1855	0.2160
Polarité	Avg-mots-possitive	0,3852(1,79%)	0,2	0.1919	0.3000
	Avg-mots-negative	0,38265(1,1%)	0,546	0.1850	0.2640
	Avg-mot-neutre	0,3751	0,53	0.1075	0.0840

FIGURE 3.2 – Impact de la pertinence dans la détection d'opinions

Nous comparons les résultats avec la base des documents pertinents données par TREC-4 nommée baseline-4. Ces dernières ont comme mesure de MAP égale à 0,3784, et de la précision égale à P@10=0,5340. Les valeurs qui contiennent un pourcentage sont celles des features qui améliorent le baseline-4.

3.2.2 Combinaison des features

À ce niveau, le but de notre travail était d'améliorer la détection d'opinions. Dans un premier temps nous avons essayé différentes combinaisons entre les différentes features pour trouver une combinaison qui donne un meilleur résultat. Dans un deuxième temps, nous devons identifier la meilleure combinaison sur différents classifieurs, afin de déterminer le meilleur classifieur et ainsi d'améliorer d'avantage la détection d'opinions (augmentation de la mesure MAP). Nous détaillons le processus de notre travail dans ce qui suit :

3.2.2.1 Meilleure combinaison

Nous avons réalisé la validation croisée 5-fois dans notre apprentissage machine. Cela signifie que cinq groupes de topics ont été créés A, B, C, D, E. Chaque groupe comporte 10 topics de TREC 2007. L'apprentissage machine comporte deux étapes. La première est la partie d'apprentissage, dans cette partie des données sont attribuées au classifieur pour l'apprentissage « training ». Ce dernier génère un modèle qui est utilisé pour la deuxième partie nommée test « testing ». L'utilisation de ces groupes dans l'apprentissage machine a été fait comme le montre le tableau 3.3.

training	testing
B, C, D, E	A
A, C, D, E	B
A, B, D, E	C
A, B, C, E	D
A, B, C, D	E

FIGURE 3.3 – Tests de validation croisée

De ce fait, un modèle est créé en utilisant le classifieur de régression logistique multinomiale [Lar06]. Les mesures utilisées sont la précision à 10 documents (P @ 10) et la précision moyenne (MAP). Les résultats montrent que la subjectivité combinée avec la réflexivité et l'adressage, y compris le niveau de pertinence du document comme est mentionné dans le tableau 3.4 peut être effective pour la recherche d'opinions dans les blogs. Cette combinaison a fait augmenter la mesure MAP de 4,51%, et de 8,61% de la précision à 10 documents. Pour prouver statistiquement l'amélioration de nos résultats la mesure t-test³ est utilisée $P < 0,05$ (0,01).

Combinaison
Emotivité-1
Subjectivité-4
Reflexivité
Addressage
Score de pertinence
Rang de pertinence

FIGURE 3.4 – Combinaison des features

3.2.2.2 l'utilisation des différents classifieurs

	MAP	P@10	clasifieur
Comb 4	0,3955	0,58	Logistic Reg
Comb 4	0,3994	0,564	SVM
Comb 4	0,3835	0,576	Naive B

FIGURE 3.5 – Résultats des différents classifieurs

3. http://www.socialresearchmethods.net/kb/stat_t.php

Le tableau 3.5 présente les résultats de la tâche de détection d'opinions en utilisant trois différents classifieurs pour la combinaison que nous avons utilisée ci-dessus. Les résultats montrent que les Support Vector Machines donnent les meilleurs résultats pour cette combinaison des features alors que le modèle de régression logistique vient en deuxième position, tandis que le naïve Bayes donne de moins bon résultats. Le résultat de 0,3994 donne une amélioration de 0,66% par rapport à la meilleur détection d'opinions qui a été faite en utilisant les topics de 2007 avec une méthode qui est basée sur la dépendance des topics, leurs MAP était égale à 0,3968 [SHMO09].

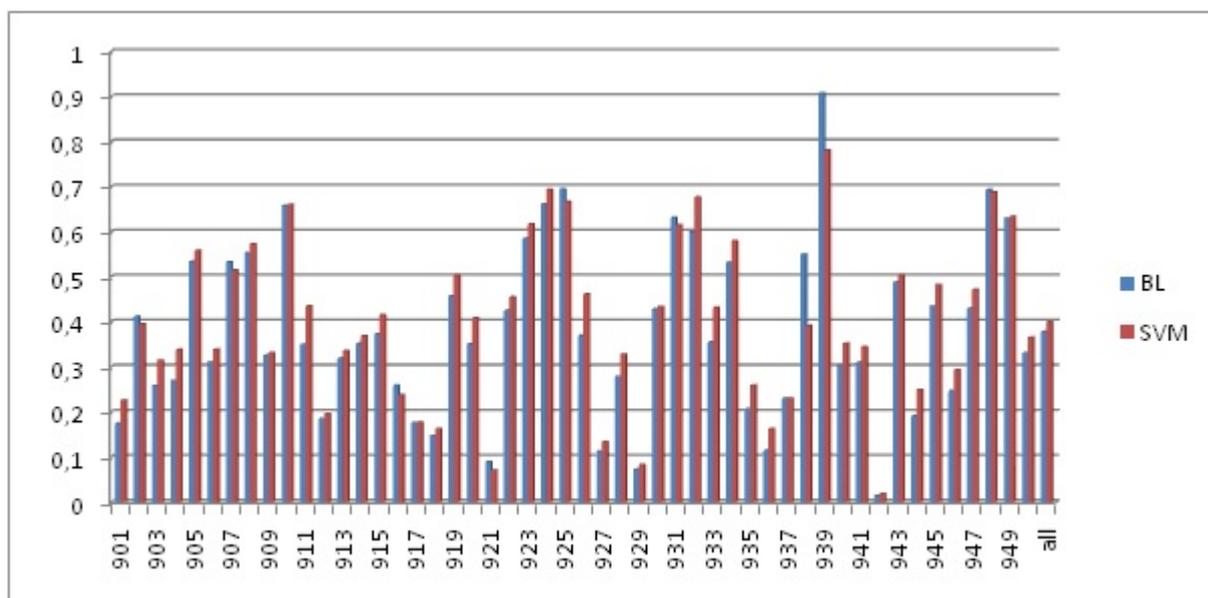


FIGURE 3.6 – L'évaluation des topics de 2007 en utilisant SVM

Le diagramme 3.6 montre le degré d'amélioration de notre run (SVM) sur base (BL) pour chaque topic de TREC 2007. Nous avons obtenu une amélioration sur plus de 40 topics sur un total de 50 topics et une amélioration moyenne de 12,26% a été observée pour 39 topics. L'amélioration maximale de 41,60% a été notée pour le topic 936 ("Grammy Awards") et l'amélioration minimale de 0,1699% a été notée par le topic 917 ("Snopes"). Le tableau 3.7 présente les 10 premières topics qui nous donnent de meilleurs résultats ainsi que leurs pourcentage d'amélioration.

Num	Topic	BL	SVM	%age Imp	Title
1	936	0.1149	0.1627	41.601	Grammy awards
2	944	0.1916	0.2496	30.271	Opera Software
3	901	0.1749	0.2255	28.930	jstor
4	904	0.2693	0.3385	25.696	alterman
5	926	0.3699	0.462	24.898	hawthorne heights
6	935	0.2073	0.2587	24.794	Mozart
7	911	0.3496	0.4351	24.456	SCI FI CHANNEL
8	903	0.2585	0.3149	21.818	Steve jobs
9	933	0.3552	0.432	21.621	winter olympics
10	946	0.2467	0.2939	19.132	tivo

FIGURE 3.7 – Topics qui donnent de bons résultats

En analysant ces résultats, nous remarquons que les topics qui ont amélioré la performance sont ceux qui concernent des sujets d'intérêt public comme le topic « Grammy Awards », le « Opera software », « Alterman » (chroniqueur et auteur), « Hawthorne Heights » (groupe musical), et « Steve Jobs », etc. Tous ces thèmes sont fréquemment utilisés dans la vie quotidienne. Dans les blogs, les gens lisent ou commentent sur les blogposts qui les intéressent. Par conséquent, la plupart des sujets d'intérêt publics obtiennent en général plus d'attention que les autres blog posts comportant des sujets moins public. Cette observation confirme également les résultats de Maarten de Rijke [Mdr06] que conclut que les blogs avec le plus grand nombre de commentaires ont tendance à contenir plus d'opinions que les blogs qui contiennent moins.

Le tableau 3.8 montre les sujets pour lesquels notre approche n'a pas amélioré les résultats. Si on regarde les titres des topics, nous constatons que les sujets ne sont pas populaires auprès du public ou que leur popularité est limitée à un certain groupe de personnes.

Num	Topic	BL	SVM	%age Imp	Title
1	902	0,4121	0,3954	-4,052	lactose gas
2	907	0,5324	0,5148	-3,305	brreeeport
3	916	0,2588	0,2375	-8,230	dice.com
4	921	0,0895	0,0706	-21,117	Christianity Today
5	925	0,6951	0,6669	-4,056	mashup camp
6	931	0,6314	0,6147	-2,644	fort mcmurray
7	938	0,5493	0,3922	-28,6	plug awards
8	939	0,9076	0,7807	-13,981	Beggin Strips
9	942	0,0143	0,0181	26,573	lawful access
10	948	0,6931	0,6879	-0,750	sorbonne

FIGURE 3.8 – Topics qui donnent les résultats les plus mauvais

On peut conclure que notre premier objectif est atteint : nous avons amélioré la détection d'opinions par rapport à la meilleure baseline de TREC (baseline-4). Dans ce qui suit nous allons introduire nos contributions par rapport à l'identification de la polarité, nous nous concentrons plutôt sur le rôle de la classification des topics dans cette tâche. Une analyse des résultats est ensuite présentée.

3.2.3 Détection de la polarité

La plupart des travaux liés à la détection de la polarité tiennent compte de la polarité (négative, positive ou neutre) des termes présents dans un document en utilisant différentes ressources lexicales. Un des problèmes dans ces travaux, est qu'ils ne prennent pas en compte les polarités contextuelles des termes. La polarité contextuelle d'un terme est la polarité qui est générée après modification de la première polarité attribuée à un terme. Cette modification de la polarité est due à un changement de contexte. Nous donnons dans ce qui suit, quelques contextes principaux responsables de changement de polarité des termes.

- la polarité attribuée au terme « bien », est positive, mais si ce terme est précédé par une négation comme « pas » ou « jamais » sa polarité change, elle devient négative. Ce changement n'aura pas pu être fait sans la prise en compte du contexte.
- la polarité contextuelle est définie par le sujet que nous recherchons. Par exemple, le mot « imprévisible » dans un document contenant des avis sur un film sera pris comme un signe positif. En revanche, si le même mot est utilisé dans un autre document par exemple pour un avis sur un appareil photo numérique et si on dit qu'une fonctionnalité de cet appareil est imprévisible alors cette fois, le mot est considéré comme un mot négatif.

Dans cette partie, nous nous intéressons à la détection de la polarité en tenant compte du contexte. Nous expérimentons notre hypothèse : la catégorisation des topics améliore les performances de cette tâche. Notre approche est mixte du fait qu'elle est basée sur le lexique SentiWordNet [ES06] et l'apprentissage machine en utilisant un modèle de régression logistique. Il est important de mentionner que notre but n'est pas d'améliorer la détection de polarité mais de voir l'impact de la classification des topics de ce fait le choix du classifieur d'import pas. Dans [ZJB07], les auteurs utilisent aussi la catégorisation des topics pour la détection de l'opinion, mais notre approche est différente sur plusieurs points de vue :

- [ZJB07] utilise une classification des topics pour la tâche de la détection d'opinions et la même classification est utilisée pour la tâche de la détection de la polarité.
- les features utilisées sont différentes.
- nous utilisons un modèle de régression logistique contrairement à [ZJB07] qui utilise les SVM.
- nous utilisons la collection de test « blog6 » proposée par TREC 2006 et 2007 avec la validation croisée dans chaque catégorie des topics, alors les paramètres expérimentaux de [ZJB07] ne sont pas clairs.

La catégorisation des topics est expliqué dans la section suivante.

3.2.3.1 Catégorisation des topics

Nous avons classé les sujets de TREC 2006 et TREC blog Blog 2007 dans 6 classes ci-dessous :

1. Film / TV ;
2. Personne ;
3. Organisation ;
4. Événement ;
5. Produit ;
6. Thème.

Cette classification est inspirée de celle des travaux de [ZJB07]. Chaque thème de blog 2006 et 2007 a été lu par deux personnes que nous avons recrutées (appelés annotateurs). Le tableau 3.9 montre l'accord des résultats entre les deux annotateurs. Dans les instructions aux annotateurs, nous leur avons demandé de lire attentivement les descriptions des sujets, puis le titre de chaque sujet et de leurs attribuer une classe parmi les classes proposées. L'étude de ce tableau a montré que l'on trouve très peu de désaccords entre les annotateurs, la plupart des désaccords se trouvent dans les sujets de l'année 2007 (15 désaccords par rapport à seulement 1 pour l'année 2006).

Annotator 1	Annotator 2						Total
	Movie/TV	Person	Org	Event	Product	Issue	
Film	12	0	0	3	0	1	16
Person	0	20	1	0	0	0	21
Org	0	0	14	3	0	2	19
Event	0	1	0	7	0	2	10
Produit	0	0	1	0	13	3	17
Theme	0	0	0	2	0	15	17
Total	12	21	16	15	13	23	100

FIGURE 3.9 – Jugement des annotateurs pour les différents thèmes

La mesure statistique κ de Cohen [Coh60] a été utilisée pour calculer le score d'accord entre les annotateurs. La mesure *Kappa* est connue pour être une mesure plus robuste que le calcul de pourcentage, car il tient compte de l'accord qui se produit par hasard. L'équation est la suivante :

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (3.10)$$

Où $Pr(a)$ est la proportion observée d'accord entre annotateurs, et $Pr(e)$ est la probabilité d'un accord hypothétique par chance. La valeur κ peut varier de -1 (totale contradiction) à 1 (plein accord). Selon notre travail la valeur κ a été calculée par l'équation (3.10), le résultat est $0,77$, ce qui représente un accord entre les deux annotateurs. Concernant les autres topics pour lesquelles les annotateurs n'ont pas été en accord, un troisième annotateur a résolu le conflit.

Le tableau 3.10 ainsi que le tableau 3.11 montrent les différentes classes attribuées a chaque topic de 2006 et 2007.

CLASS	TOPIC 2006	TOTAL
EVENT	853 861 890	3
THEME	855 ,858 ,859 ,865 ,867,868,869,878,889,894,896,898, 899	13
TV	851, 860, 864, 872, 875,876, 881, 886, 895	9
PERS	852, 854, 857, 870, 871, 873, 874, 880, 891, 892, 897	11
PROD	856, 862, 879, 883, 893	5
ORG	863, 866, 877, 882, 884, 885, 887, 888, 900	9

FIGURE 3.10 – Classification en thèmes des topics de blog 2006

CLASS	TOPIC 2007	TOTAL
EVENT	943, 938, 936, 933, 925, 923, 914, 913, 906, 905	10
THEME	901, 902, 907, 918, 921, 927, 929, 931, 942	9
TV	911, 922, 928	3
PERS	903, 904, 908, 920, 924, 935, 940, 941, 947, 949	10
PROD	909, 916, 917, 932, 934, 937, 939, 944, 946, 950	10
ORG	910, 912, 915, 919, 926, 930, 945, 948	8

FIGURE 3.11 – Classification en thèmes des topics de blog 2007

Dans ce qui suit, nous effectuons des expériences sur la détection de polarité en utilisant cette catégorisation des topics. Nous allons ensuite analyser les effets de cette catégorisation. Pour cela nous effectuons des expériences en deux phases. Dans la première phase, nous réalisons des expériences sur la détection de polarité sans classification des topics et dans la seconde phase, nous re-expérimentons avec des topics de TREC 2006 et 2007 en utilisant la classification des topics .

3.2.4 Détection de polarité sans classification des topics

Nous utilisons un modèle de régression logistique pour nos expériences. Les features utilisées pour détecter la polarité des documents sont présentés dans le tableau 3.12.

Features	Description
Total-Pos	Avg des mots positives dans D
Total-Neg	Avg des mots negatives dans D
Total-Net	Avg des mots neutres dans D
Total-Adj	Avg des adjectives dans D

FIGURE 3.12 – Les features de la polarité

Nous avons choisi quelques features simples et ordinaires (déjà utilisés dans [LH08]) afin d'éviter toute complexité. Il est à rappeler que ce travail ne vise pas à proposer de nouvelles fonctionnalités permettant de détecter la polarité des documents, mais de vérifier l'impact de la classification des topics dans la polarité. Pour déterminer la polarité, il nous a fallu une base de départ avec des documents qui contiennent des opinions. Pour cela nous utilisons le run donné par la feature subjectivité-4 (dont la MAP est de 0,3879 et la précision P@10 de 0,546) comme point de départ. Les expérimentations pour la détection de polarité sans classification ont été effectuées dans trois environnements différents. Toutes les expériences avec leurs paramètres sont expliquées ci-dessous.

1. *expérimentation 1*

Les expériences ont été effectuées en utilisant les mêmes features que ceux expliqués ci-dessus. Une validation croisée à 5 a été réalisée pour les sujets de l'année 2007. Le résultat pour la mesure MAP Positive est de 0,0999 et pour la mesure de précision p@10 de 0,2000 tandis que pour la mesure MAP Négative, le résultat est égal à 0,0651 et pour P@10, il est égal à 0,0609.

Problème avec cette expérimentation

Dans cette expérimentation le volume de données utilisées dans la partie d'apprentissage est beaucoup plus grand que celui des données utilisées pour la classification. Par conséquent, avant de discuter d'autres causes qui pourraient améliorer ces résultats, nous procédons à une autre expérience en utilisant un nombre réduit de données d'entraînement pour des expériences sans classification de topics.

2. *expérimentation 2*

Dans ce contexte, les données d'apprentissage ont été ramenées de 40 à 22 topics. 22 est le nombre maximum de sujets d'un groupe dans l'expérience avec classification de topics, par conséquent, nous avons réduit le nombre de sujets à 22 pour l'apprentissage. Le résultat obtenu pour la MAP Positive est de 0,1636 et pour la précision p@10 de 0,2000 tandis que la MAP Négative est égale à 0,0629 et la P@10 est de 0,0587.

Problème avec cette expérimentation

Le choix des sujets d'essai pour cette expérience a été fait de manière aléatoire. Le choix des topics d'essai a été fait dans l'ordre numérique : le premier essai a été fait pour les sujets de 901 à 910, le deuxième pour les sujets de 911 à 920 tandis que les topics d'apprentissage, ils ont été choisis au hasard. Le problème qui peut donc se poser est que des topics de la même classe peuvent être dans la partie test et dans la partie apprentissage, ce qui devrait être évité. Par conséquent, nous effectuons une autre expérimentation en utilisant un troisième paramètre.

3. *expérimentation 3*

Pour résoudre le problème de la deuxième expérimentation et améliorer davantage la détection de la polarité, nous avons différencié les topics pour le test et les topics pour l'apprentissage. Pour les données de test, nous utilisons l'ensemble des sujets de la classe « Event » (ou de la classe « produit » ou d'une autre classe) et pour les données d'apprentissage, nous utilisons 22 autres sujets qui n'appartiennent pas à la classe relative aux données d'essai. Ce paramètre a également donné de bons résultats pour l'expérimentation avec classification des topics mais une grosse chute dans la MAP a été constatée. Il est donc important de souligner que la présence des topics de la même classe dans la partie apprentissage et dans la partie test a un impact considérable sur les performances. Le résultat est pour la MAP Positive MAP de 0,0557, et pour la précision @10 de 0,072 tandis que la MAP Négative est égale à 0,0369 et la mesure de P@10 est égale à 0,0543. Le tableau 3.13 récapitule les résultats des trois expérimentations

	POS		NEG	
	MAP	P@10	MAP	P@10
Expérimentation 1	0,099	0,2	0,0651	0,0609
Expérimentation 2	0,1636	0,2	0,0629	0,0587
Expérimentation 3	0,0557	0,072	0,0369	0,0543

FIGURE 3.13 – Résultats des différentes expérimentations de la polarité

3.2.5 Détection de la polarité avec classification des topics

Dans cette partie nous utilisons les mêmes features que ceux utilisées pour la détection de la polarité sans classification, c'est-à-dire : nombre de mots positifs, nombre de mots négatifs, nombres de mots neutres, ainsi que le nombre d'adjectifs. Un groupe de topics a été créé pour chaque classe. Nous avons considéré les topics de TREC 2006 et TREC 2007. Une validation croisée à n a été réalisée parmi les sujets de chaque groupe, en utilisant un modèle de régression logistique, où n est le nombre de sujets dans un groupe. Nous nous sommes reportés uniquement aux résultats de TREC 2007. Les groupes définis pour chaque classe sont donnés dans le tableau 3.14.

CLASS	TOPIC 2006	TOPIC 2007	TOTAL
EVENT	853,861,890 = 3	943, 938, 936, 933, 925, 923, 914, 913, 906, 905 = 10	13
THEME	855 ,858 ,859 ,865 ,867,868,869,878,889,894,896,898, 899 = 13	901, 902, 907, 918, 921, 927, 929, 931, 942 = 9	22
TV	851,860,864,872,875, 876, 881,886,895 = 9	911, 922, 928 = 3	12
PERS	852, 854, 857, 870, 871, 873, 874, 880, 891, 892, 897= 11	903, 904, 908, 920, 924, 935, 940, 941, 947, 949= 10	21
PROD	856, 862, 879, 883, 893 = 5	909, 916, 917, 932, 934, 937, 939, 944, 946, 950= 10	15
ORG	863, 866, 877, 882, 884, 885, 887, 888, 900= 9	910, 912, 915, 919, 926, 930, 945, 948 = 8	17

FIGURE 3.14 – Les groupes de la classification des topics

Nous avons inclus les sujets de TREC 2006, afin d'augmenter les données d'entraînement pour l'apprentissage machine de sorte que la comparaison entre « avec classification des topics » et « sans classification des topics » peut être justifiée. Les résultats de cette expérimentation sont donnés dans le tableau 3.15

	POS		NEG	
	MAP	P@10	MAP	P@10
Expérimentation 3	0,055	0,072	0,036	0,054
Sans classification	0,109	0,146	0,068	0,068
% amélioration	9688,30%	102,77%	85,24%	26,87%

FIGURE 3.15 – Comparaison des résultats de la polarité

Ce tableau présente les résultats de la détection de polarité selon la classification des topics pour les polarités positives et négatives. Ces résultats montrent que la classification des topics a bien amélioré les résultats pour toutes les expérimentations qu'on a faites. Nous nous reportant juste par rapport à la dernière expérience qui est considérée comme la plus adaptée à notre travail. Une amélioration considérable peut être notée dans les résultats, cela est du au choix de notre baseline qui est très faible. Toutefois, ces résultats sont suffisants pour prouver que la catégorisation des topics peut améliorer les résultats de la détection de la polarité. Il est à noter que l'objectif de ce travail n'était pas d'améliorer le travail précédent pour la détection de la polarité, mais plutôt d'analyser les effets de la classification sur la tâche de détection d'opinions.

Les diagrammes 3.16 et 3.17 ci-dessous montrent une amélioration de la mesure MAP de chaque topic de TREC 2007 pour les polarités positives et négatives.

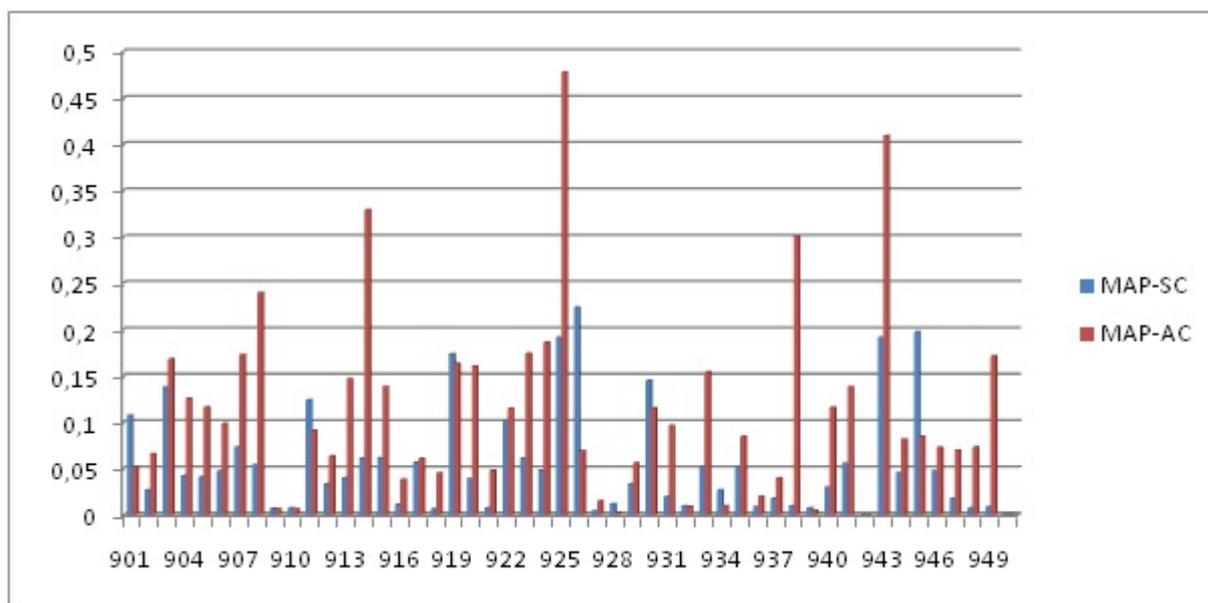


FIGURE 3.16 – Polarité positive

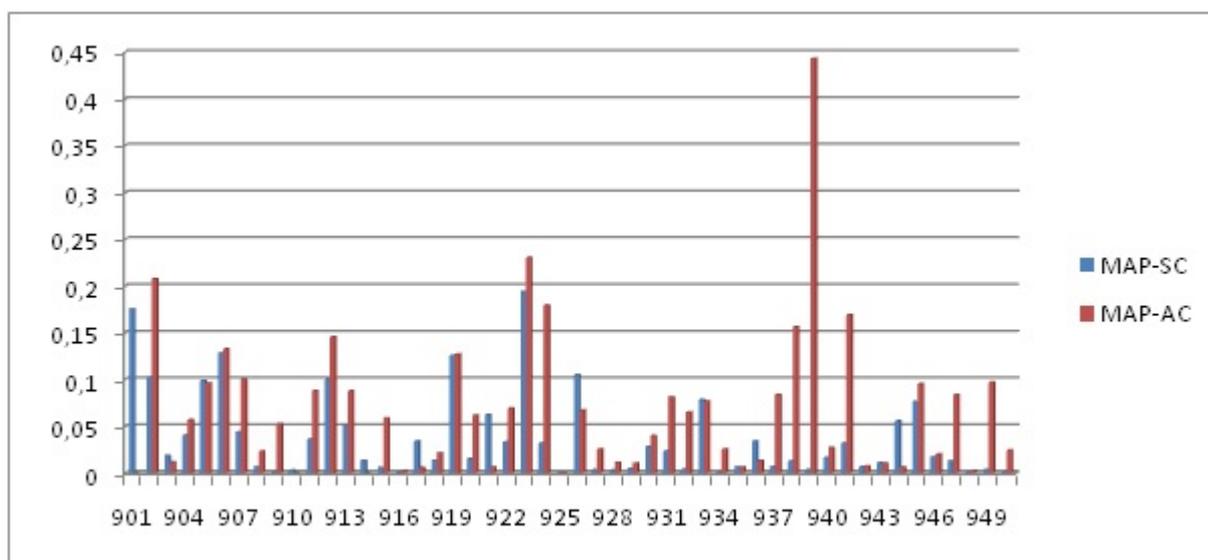


FIGURE 3.17 – Polarité négative

En étudiant les diagrammes de la polarité positive et de la polarité négative, on note que la plupart des topics pour lesquels une amélioration importante a été constaté dans les deux polarités, sont ceux qui appartiennent aux classes « Event », « Thème » ou « Person » (comme 902, 907, 908, 924, 938, etc.). L'une des raisons pour laquelle une importante amélioration est obtenue

dans ces classes est peut être du au nombre de topics dans les jeux de données d'apprentissage. La classe « Thème » et la classe « Personne » ont un nombre de topics de 22 et 21, respectivement dans leur ensemble. Cependant, cette justification ne tient pas pour la classe « événements » où nous avons 13 topics au total qui est inférieur à la classe « ORG » (17 topics) et à la classe « PROD » (15). Une raison possible pourrait être la classification elle-même des topics qu'on a faite. Nous avons observé que la plupart des conflits rencontrés lors de la classification des topics était de devoir décider entre les topics classifiés en tant qu' « événement » et les topics classifiés en tant que « Thème ». Par exemple, il était difficile de décider si le « discours » du président est un thème ou un événement ? Par conséquent, nous supposons que des améliorations dans les sujets de la classe « événement » pourraient également refléter des conflits de processus de classification des sujets.

Ce que nous avons fait dans un dernier temps est d'introduire la classification au niveau de la détection d'opinions. Nous avons constaté que la classification des topics n'a pas amélioré les résultats de la détection d'opinions.

Le tableau 3.18 montre les différents résultats de chaque expérimentation. Cependant, le résultat de la troisième expérimentation (sans classification) nous permet de déduire que la classification des topics impacte sur le résultat de la détection d'opinions. Dans la troisième expérimentation (sans classification), nous avons fait apprendre nos données avec 22 topics en séparant les topics de la même classe. Cela veut dire, que les topics lors de l'étape d'apprentissage diffèrent de ceux lors de l'étape des tests. Cette méthode, nous a permis de comparer les deux méthodes « la classification » et « sans classification ». La réduction de la MAP de 0,3934 à 0,3656 indique que la présence des topics de la même classe dans la partie apprentissage ainsi que dans la partie de teste est bénéfique.

Nous concluons que nos résultats confirment les résultats obtenus par Z et al. [ZJB07], où le classement des topics pour la détection d'opinions pourrait simplement améliorer une seule baseline mais ne pourrait pas fonctionner correctement sur la deuxième baseline. Par conséquent, nous pensons encore qu'une expérimentation rigoureuse est nécessaire à cet égard. Nous la réaliserons dans les travaux futurs.

Classification	Expérience	Mesures	Résultats
Sans classification	1er	MAP	0,3955
		P@10	0,58
	2em	MAP	0,3934
		P@10	0,58
	3em	MAP	0,3656
		P@10	0,534
Avec classification		MAP	0,3296(-9,84)
		P@10	0,534

FIGURE 3.18 – Détection d'opinions avec classification

3.2.6 Bilan

Dans ce chapitre, nous avons proposé une approche original pour la détection de l'opinion. Nous avons vu que notre méthode améliore non seulement la baseline, mais aussi les résultats rapportés précédemment pour les topics de TREC 2007. En outre, nous avons également analysé les effets de la classification des topics sur la tâche de la détection de la polarité. Les résultats montrent que le classification des topics améliore les résultats pour la détection de la polarité. Nous avons essayé cette classification pour la détection d'opinions, les résultats ne sont pas satisfaisants. Dans notre travail le classement des topics a été réalisé manuellement, nous proposons dans le futur d'utiliser des algorithmes de classification d'apprentissage machine et des services d'annuaire populaires comme Yahoo⁴ ou Dmoz⁵. En outre, nous avons l'intention de combiner notre approche qui est basée sur l'indépendante du topic avec une approche qui dépend du topic et voir si les résultats peuvent s'améliorer.

4. <http://www.yahoo.fr/>

5. <http://www.Dmoz.org/>

Conclusion et perspectives

Conclusion

Le web fournit une importante collection d'opinions sous forme de blogosphère. Les chercheurs travaillant dans le domaine de la détection d'opinions pensent que cette nouvelle collection de données sous forme de blogs peut donner de nouvelles directions dans ce domaine. Vu l'importance de cette collection de données, TREC (Text retrieval conference) a proposé au blog track en 2006 avec une collection de données appelée collection TREC Blog 2006 (de 148 Go). De 2006 à 2008, chaque année 50 nouveaux topics ont été proposés pour expérimentation. Cette collection a été fournie avec les « qrels ». La détection d'opinions a été la principale tâche de TREC Blog 2006. Cependant en 2007, une nouvelle sous tâche a été proposée. Elle consiste à détecter la polarité d'opinions. Dans ce rapport, nous avons présenté notre travail sur la détection d'opinions et sur la détection de la polarité :

- une proposition d'une approche pour la détection d'opinions indépendante des topics ;
- une classification des topics de TREC Blog 2006 et 2007 ;
- une expérimentation de la détection de la polarité sans classification des topics ;
- une expérimentation de la détection de la polarité avec classification des topics ;
- Une expérimentation de la détection d'opinions sans classification des topics ;
- une expérimentation de la détection d'opinions avec classification des topics.

À la vu de notre expérimentation et des résultats obtenus, il s'avère que :

- notre approche de détection d'opinions a amélioré la meilleur « baseline » de TREC. En utilisant l'algorithme d'apprentissage machine de type SVM, les résultats ont donné une MAP de 0,3994 comparée à 0,3968 soit 0,66% d'amélioration ;
- la classification des topics a été faite par deux annotateurs ;
Tous les topics de TREC 2006 et 2007 ont été classifiés dans l'une des 6 classes que nous avons définies : (« event », « thème », « organisation », « produit », « person », « TV »). La valeur de 0,77 de Kappa a été trouvée et donne une mesure de la ressemblance de la décision entre les deux annotateurs. Cette valeur indique un bon accord entre les deux annotateurs ;
- nous avons observé que la classification des topics améliore les résultats de la détection de la polarité. Les résultats avec classifications sont meilleurs que ceux effectués sans classi-

fication ;

- pour la détection d’opinions, la classification n’améliore pas les résultats. Cependant, les résultats donnent une indication sur le rôle de la classification des topics dans cette tâche.

Les résultats obtenus lors de nos expérimentations sont probants. Un ensemble de problèmes restent cependant à résoudre.

- la classification des topics a été faite manuellement dans notre travail. Nous proposons dans le futur une classification automatique des topics. Pour cela, on peut tirer profit des services d’annuaires populaires tels que yahoo, ou DMoz ;
- nous planifions aussi de combiner notre approche de la recherche d’opinions indépendante des topics avec une recherche d’opinions basée sur les topics (comme la proximité entre le titre du topic et les termes d’opinion), pour améliorer les résultats.

Bibliographie

- [ABM09] Asher N., Benamara F. et Mathieu Y. : Appraisal of Opinion Expressions in Discourse. *Linguisticæ Investigationes*, 32:2, 2009.
- [CCC⁺08] Cabanac G., Chevalier M., Chrisment C., Julien C., Soulé-Dupuy C. et Tchienehom P. : Web Information Retrieval: Towards Social Information Search Assistants. Dans Kidd T. et Chen I., éditeurs : *Social Information Technology: Connecting Society and Cultural Issues*, chapitre 16, p. 218–252. IGI Global, mars 2008.
- [CCCJ10] Cabanac G., Chevalier M., Chrisment C. et Julien C. : Social validation of collective annotations: Definition and experiment. *J. Am. Soc. Inf. Sci. Technol.*, 61(2):271–287, feb 2010.
- Cet article est basé sur le calcul de la validation sociale dans les annotations collectives. Le but est de savoir l’avi collectifs d’un ensemble d’annoteurs. Trois algorithmes sont proposés : l’algorithme 1 basé sur le coefficient de Cohen kappa. L’algorithme 2 empirique récursif (agrement d’une annotation et combinaison de ces agrements). L’algorithme 3 est plus formel basé sur le modele de Dung crée un arbre <A,B> ou A est l’ensemble des annotations et B est la relation entre les annotations (confirme, refute). Les tests ont été faits à partir de données récoltées en ligne de 121 personnes. Les trois algorithmes ont donné des résultats similaires de 84% par rapport à la perception humaine.
- [CHBC10] Cabanac G., Hubert G., Boughanem M. et Chrisment C. : Impact du « biais des *ex aequo* » dans les évaluations de Recherche d’Information. Dans *CORIA’10 : Actes de la 7^e conférence en recherche d’information et applications*, p. 83–98, mars 2010.
- Au sein de l’évaluation de la recherche d’information , deux problématiques se posent au niveaux des mesures utilisées par TREC -EVAL : 1 la stratégie du bourrage (ajout de documents non pertinents pour atteindre le nombre 1000) ce qui affecte le résultat. 2 le traitement des documents équivalents (même similarité). TREC-Eval ordonne les documents de même similarité par ordre alphabétique décroissant. Ceci favorise certains documents par rapport à d’autres (en affectant des rangs différents). Deux méthodes on été proposées par les auteurs. 1 méthode réaliste : Les documents de même similarité sont ordonnés selon la pertinence les documents non pertinents sont mis avant les pertinents. Si les documents non pertinents (ou pertinents) sont équitables alors l’ordre alphabétique décroissant sera appliqué. 2 méthode optimiste : même procédure sauf que les documents pertinents sont classés avant les non pertinents. Les mesures utilisées sont : RR RP MAP. La comparaisent se portera sur la méthode réaliste et conventionnelle (celle de TRECEval).Les résultats de la comparaison montrent que les deux méthodes sont corrélées, à plus de 87% mais assez différentes pour faire changer le résultat. Les auteurs proposent d’intégrer ces deux méthodes dans TRECEval.
- [Coh60] Cohen J. : A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20:37–46, 1960.
- [Coh06] Cohen J., éditeur. *A Coefficient of Agreement for Nominal Scales*. Educ. Psychol.Meas, 2006.
- [DF04] Drezner D. et Farrell H. : The power and politics of blogs. 2004.

- [DLP03] Dave K., Lawrence S. et Pennock D. M. : Mining the peanut gallery: opinion extraction and semantic classification of product reviews. p. 519–528, 2003.
- [ES06] Esuli A. et Sebastiani F., éditeurs. *SentiWordNet: A publicly available lexical resource for opinion mining*. LREC-06: in Proceedings of Language Resources and Evaluation Conference, European Language Resources Association, Genova, 2006.
- [Esu06] Esuli A. : Opinion mining. 2006.
- [FS08] Fautsch C. et Savoy J. : UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogosphere. *Dans TREC: Proceedings of the Text Retrieval Conference*, 2008.
- [HMP⁺07] Hannah D., Macdonald C., Peng J., He B. et Ounis I. : University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier. *Dans TREC: Proceedings of the Text Retrieval Conference*, 2007.
Les auteurs utilisent deux approches pour détecter les opinions dans les blogs. La première approche est basée sur un dictionnaire de 12 000 mots construit à partir du corpus. La deuxième approche utilise Opinion Finder. Les résultats montrent que la première approche prend moins de temps que la deuxième, et qu'elle améliore la MAP de 15,8% par rapport au « baseline ». Cette approche est utilisée aussi pour le calcul de la polarité
- [HT09] H Tang, S Tan X. c., éditeur. *A survey on sentiment detection of reviews*, volume Special Publication 500-277. Journal Expert System with Application 36, 2009.
Les auteurs présentent les différentes approches pour le problème de la détection des sentiments dans les textes. Ils divisent ce dernier en quatre sous problèmes liés entre eux : classification subjective, classification de sentiment du mot, classification de sentiment du document et extraction de l'opinion.
- [JWR00] Jones K. S., Walker S. et Robertson S. E. : A probabilistic model of information retrieval: development and comparative experiments - part 1. *Inf. Process. Manage.*, 36(6):779–808, 2000.
- [KS05] Koppel M. et Schler J. : Using neutral examples for learning polarity (short paper). *International Joint Conferences on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, July 2005.
- [Lar06] Larose D. T., éditeur. *Data Mining Methods and Models*, volume ISBN 0-471-66656-4 344p. New York: Wiley, 2006.
- [LCW⁺07] Liao X., Cao D., Wang Y., Liu W., Tan S., Xu H. et Cheng X. : Experiments in trec 2007 blog opinion task at cas-ict. *Dans TREC*, 2007.
Cet article présente les travaux de « l'Institute of Computing Technology of Beijing of China » sur la détection de l'opinion et la polarité dans Trec 2007. Les auteurs s'intéressent au blog entier (blog post et ses commentaires) et utilisent le classifieur Drag-Push pour la détection de l'opinion et la polarité. Ils utilisent aussi les SVM's pour filtrer les blogs qui sont des spams.
- [LH08] Linh Hoang, Seung-Wook Lee G. H. J.-Y. L. H.-C. R., éditeur. *A Hybrid Method for Opinion Finding Task (KUNLP at TREC 2008 Blog Track)*, 2008.
- [Lia]
- [Liu07] Liu B. : *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer, 2007.
Ce chapitre utilise trois aspects évaluatifs d'un texte. 1 classification des sentiments : détermine la polarité pour tout le document, les applications utilisées pour cet aspect la sont : *classification basée sur les phrases sentimentales* (utilisation de Part-of-speech, extraction des adjectifs et des adverbes, calcule la sémantique opinion entre les phrases et les mots pour déterminer une polarité), *classification d'un texte* (détermine une classe positive ou négative pour un texte, utilise pour cela naïve Bayesian ou SVM ou autre), *utilisation basée sur un score* (détermine pour chaque terme

une classe, donne un score a ses termes et classe le document). 2 Featured based : determine la polarité au niveau de la phrase, extrait toutes les différents features et synonymes de l'objet. 3 comparaison des phrases : l'importance de cette approche est de pouvoir mesurer l'importance d'un objet comparativement a un autre. Le chapitre presente de nombreux algorithmes mais l'évaluation et leur comparaison n'est pas abordée

- [LNK⁺08] Lee Y., Na S.-H., Kim J., Nam S.-H., Jung H.-Y. et Lee J.-H. : Kle at trec 2008 blog track: Blog post and feed retrieval. *Dans TREC*, 2008.
- Cet article présente l'approche proposée par l'université de Pohang (Republic of Korea) à TREC 2008 pour la recherche d'opinions et la polarité. Pour la détection d'opinions, Les auteurs proposent une approche basée sur le lexique. Ce dernier est créé en utilisant Sentiwordnet et « Amazon's review corpus ». La requête initiale est étendue par un ensemble représentatif de tous les mots subjectifs. La même approche est utilisée pour la polarité sauf qu'au lieu de considérer la subjectivité du mot, ils considèrent l'orientation sémantique (polarité), et la requête est constituée seulement du titre du topic.
- [MB09] Missen M. M. S. et Boughanem M. : Sentence-level opinion-topic association for opinion detection in blogs. *Dans AINA Workshops*, p. 733–737. IEEE Computer Society, 2009.
- [MBC09a] Missen M. M. S., Boughanem M. et Cabanac G. : Challenges for Sentence Level Opinion Detection in Blogs. *Dans ICIS '09: Proceedings of the 2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*, p. 347–351, Washington, DC, USA, 2009. IEEE Computer Society.
- Cet article traite la détection d'opinion au niveau des phrases. Le problème qui se pose dans ce cas est la difficulté d'extraire les phrases dans le texte car la ponctuation n'est pas respectée et le langage utilisé dans les blogs est particulier. La perception de la polarité diffère selon les annotateurs et la métrique utilisée dans le document (somme d'opinions positives et somme d'opinions négatives) n'est pas adaptée au niveau de la phrase. La solution proposée pour mesurer la divergence d'opinion est de calculer une mesure statistique de κ Cohen. Cette mesure permet d'observer la divergence ou la convergence des avis de différents internautes.
- [MBC09b] Missen M. M. S., Boughanem M. et Cabanac G. : Comparing semantic associations in sentences and paragraphs for opinion detection in blogs. *Dans MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, p. 483–488, New York, NY, USA, 2009. ACM.
- Les auteurs comparent la détection d'opinion au niveau d'une phrase avec la détection d'opinion au niveau d'un paragraphe et propose trois approches. La première approche concerne la phrase et utilise *OKAPI*. Elle combine le score d'opinion selon six composantes : subjectivité, émotivité, réflexivité, addressible, phrase, OTA. OTA utilise Wordnet pour désambiguïser et ajouter de la sémantique. De plus, la requête est étendue en utilisant Wikipédia. La deuxième approche est une amélioration de la première en éliminant le score de l'émotivité et en calculant OTA juste pour les phrases subjectives qui contiennent plus d'un adjectif. La troisième approche s'est focalisée sur les paragraphes en identifiant, et en enlevant les mots inutiles et en sélectionnant les passages pertinents. Le meilleur résultat a été donné par la troisième méthode qui améliore la baseline de 32,26%.
- [MC08] Macdonald C, Ounis e. S., éditeur. *Overview of the TREC 2007 Blog Track. In Proceedings of the TREC 2007*, 2008.
- La conférence de TREC 2007 s'est spécialisée dans deux taches, la détection de l'opinion et polarité et la distillation des blogs. La collection utilisée est celle de TREC 2006 avec 50 topics en plus des 50 de TREC2006. Les mesures utilisées sont MAP, Rprec et p@10 et pour la polarité r-accuracy, A@10 et A@1000. Il y a eu 40 participants, les meilleurs résultats de la détection de polarité ont été obtenus par l'Université de Glasgow, l'Université de Indiana, l'Université d'Arkansas de Little Rock et l'Université de Waterloo.

- [MdR06] Mishne G. et de Rijke M. : A study of blog search. *Dans ECIR*, p. 289–301, 2006.
- [MG10] Macdonald G O. e. S., éditeur. *Overview of the TREC 2009 Blog Track In Proceedings of the TREC 2009*, 2010.
Cette conférence s’est intéressée aux tâches de « faceted blog distillation » et « top stories identification task ». Une nouvelle collection dite blog08 a été utilisée, et qui a été collectée de janvier 2008 à février 2009. Neuf groupes ont participé pour la première tâche et sept groupes pour la seconde.
- [MOS07] Macdonald C., Ounis I. et Soboroff I. : Overview of the TREC 2007 Blog Track. *Dans TREC:Proceedings of the Text Retrieval Conference*, 2007.
TREC 2007 a été une extension de TREC 2006. Il s’est spécialisé dans deux tâches. La première est la détection de la polarité et la deuxième est la *distillation* des blogs. La collection utilisée est celle de TREC 2006 avec 50 requêtes. Les mesures utilisées par TREC sont MAP Rprec, p@10 et pour la polarité les mesures utilisées sont r-accuracy, A@10 et A@1000. Il y a eu 40 participants, les meilleurs résultats de la détection de polarité ont été obtenus par l’Université de Glasgow, Université de Indiana, Université d’Arkansas de Little Rock et Université de Waterloo.
- [OMS08] Ounis I., Macdonald C. et Soboroff I. : Overview of TREC-2008 Blog Track. *Dans TREC:Proceedings of the Text Retrieval Conference*, 2008.
La conférence TREC2008 s’est intéressée aux tâches de recherche de blog post pertinents, de détection d’opinion et de polarité et de distillation de blogs. 50 autres topics ont été rajoutés aux 100 de TREC2007. Vingt groupes ont participé à cette conférence (vingt pour le baseline, dix neuf pour la détection d’opinion, seize pour la polarité et douze pour la distillation de blog.
- [Opi05] *Opin, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*. The Association for Computational Linguistics, 2005.
Opinion finder est un système qui permet de détecter automatiquement les phrases subjectifs (qui contiennent une opinion), il permet aussi de faire une lemmatisation des mots, et ajoute des étiquettes aux mots (dire si le mot est un nom, verbe, adjectif...). Il a été utilisé dans plusieurs travaux de détection d’opinions.
- [OZ08] O Zubaryeva J. S. : Opinion and Polarity Detection within Far-East Languages in NTCIR-7. 2008.
Les travaux présentés à NTCIR-7 par l’université de Neuchatel (Switzerland) sont exposés dans cet article. Les auteurs s’intéressent l’opinion et la polarité au niveau d’une phrase. L’approche utilisée est basée sur l’apprentissage machine de type « logistic regression ».
- [OZ10] O Zubaryeva J. S. : Evaluation de modèles de classification appliqués a la détection d’opinion. *Dans Conférence francophone en Recherche d’Information et Applications (CORIA), Sousse, Tunisie*, 2010.
Cet article parle de la détection d’opinion au niveau de la phrase. La collection utilisée est de NTCIR 2006-2007 (comportant des extraits d’articles Anglais, Chinois et Japonais). Les documents pertinents sont connus, ils sont jugés par trois personnes différentes. La collection comporte alors 10145 phrases (24,6% qui comporte une opinion et 75,4% ne comportant pas d’opinion). Les mesures utilisées pour l’évaluation sont : le rappel , la précision et une nouvelle mesure de performance basée sur la précision . La problématique posée est que la détection d’opinion au niveau des phrases est difficile a déterminer que ce soit pour la méthode basé sur le lexique (levin 1993, catégories verba, négation, fréquences des coocurrences) ou pour l’apprentissage (la représentation des documents ou des phrases , les différents modèles a utilisés). La nouvelle approche qui a été proposée (apprentissage) est basée sur le calcul d’un score Z. La comparaison a été faite entre la méthode SVM, Naive Bayes et le score Z. Le meilleur résultat a été donné par la méthode du score Z elle est plus performante de plus de 21% par rapport a SVM et de plus de 46% que Naive Bayes

- [PL08] Pang B. et Lee L., éditeurs. *Opinion Mining and sentiment analysis Foundations and Trends in Information Retrieval*, 2008.
Cet article présente les techniques et les approches pour la recherche d'opinion et l'analyse du sentiment. Les campagnes d'évaluation et les « Benchmark datasets » sont aussi présentés.
- [PLV02] Pang B., Lee L. et Vaithyanathan S. : Thumbs up?: Sentiment Classification using Machine Learning Techniques. *Dans EMNLP'02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, p. 79–86, Morristown, NJ, USA, 2002. ACL.
Cet article compare les techniques de l'apprentissage machine pour la classification des sentiments
- [SHMO09] Santos R. L. T., He B., Macdonald C. et Ounis I. : Integrating proximity to subjective sentences for blog opinion retrieval. *Dans ECIR*, p. 325–336, 2009.
- [STS+07] Song R., Tang Q., Shi D., Lin H. et Yang Z. : DUTIR at TREC 2007 Blog Track. *Dans TREC:Proceedings of the Text Retrieval Conference*, 2007.
L'article présente les approches proposées par l'université de Beiyong de chine pour les taches de recherche d'opinion et polarité et le Feed distillation. Pour la détection d'opinion, les auteurs utilisent une approche basée sur un lexique de 2000 mots subjectifs construit à partir du corpus. La requête est étendue par des mots du descriptif, du narratif et de wikipedia. Pour le calcul d'opinion, les auteurs utilisent une méthode basée sur le Gain d'Information et les SVM.
- [VB07] Voorhees E. M. et Buckland L. P., éditeurs. *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.
- [VB08] Voorhees E. M. et Buckland L. P., éditeurs. *Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST), 2008.
- [Vec07] Vechtomova O. : Using subjective adjectives in opinion retrieval from blogs. *Dans TREC:Proceedings of the Text Retrieval Conference*, 2007.
Cet article, pour la detection d'opinion utilise la ressource(Hatzivassiloglou and McKeown de 1336 subjectifs adjectifs), et implemente plusieurs approches. La premiere nommée UW opinion1 utilise pour le *baseline* un matching terme par terme et pour la detection d'opinion de meme. La seconde pour le *baseline* un matching se fait au niveau du terme et pour l'opinion au niveau de la phrase. La troisième pour le *baseline* au niveau de la phrase et au niveau du terme pour l'opinion. La quatrieme pour le *baseline* au niveau de la phrase et de meme pour l'opinion. La meilleure MAP a été donnée par la troisieme approche elle est de 0,349.
- [YH03] Yu H. et Hatzivassiloglou V. : Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sapporo, Japan, July 2003.
- [YS07] Yohei Seki, D Lun-wei Ku H. C.-N. k. C. L. : Overview of Opinion Analysis Pilot tas kat NTCIR-6. *Dans NTCIR-6 Workshop Meeting, Tokyo, Japan*, 2007.
- [YYZ07] Yang K., Yu N. et Zhang H. : WIDIT in TREC 2007 Blog Track: Combining Lexicon-Based Methods to Detect Opinionated Blogs. *Dans TREC: Proceedings of the Text Retrival Conference*, 2007.
Cet article présente l'approche proposée par « Indiana University » à TREC 2007 pour la recherche d'opinions et la polarité. Cette approche est basée sur cinq lexiques : HF lexique (High Frequency)

qui contient les termes subjectifs les plus couramment utilisés, WL lexique (Wilson's Lexique), LF lexique (Low Frequency) qui correspond aux mots créés par les gens pour exprimer leur opinion, IU lexique (I an You) qui est constitué par les n-grammes commençant ou se terminant par un IU terme et enfin le OA lexique (Opinion Acronym) qui contient les sigles représentant une opinion. Le score d'opinion d'un document est calculé selon une pondération des scores des différents lexiques. Les résultats sont comparés avec ceux du baseline. L'approche améliore la mesure MAP de 15% pour les requêtes courtes et de 11% pour les requêtes longues.

[ZJB07] Zhou G., Joshi H. et Bayrak C. : Topic categorization for relevancy and opinion detection. *Dans TREC:Proceedings of the Text Retrieval Conference, 2007.*

Cet article propose trois approches pour la détection d'opinion dans les blogs. La première approche considère que les mots ('i', 'me', 'you', 'we', 'like', 'feel', 'think' et autres) représentent une opinion. Elle utilise le moteur de recherche INDRI, pour calculer le nombre de ces mots dans une fenêtre contenant 20 mots. La deuxième approche est basée sur l'apprentissage machine et utilise le classifieur SVM. Les topics sont répartis en six catégories. La troisième approche est basée sur le traitement du langage naturel et traite la pertinence et la détection d'opinion en une seule étape. Le meilleur résultat a été donné par la première approche (sans extension de la requête). Elle améliore de 3,7% le « baseline », de 2,6% avec extension de la requête, de 3,9% la deuxième approche et de 8% la troisième approche.

Résumé

Ce document décrit notre travail sur la détection d'opinions et de la polarité dans les blogs. L'approche proposée pour la détection de l'opinion est indépendante des topics et est basée sur le lexique SentiWordNet et l'apprentissage machine. Concernant la polarité, nous avons classifié les topics en six classes, et nous avons étudié l'impact de cette classification sur la détection de la polarité.

Mots-clés: Recherche d'information, opinions, polarité

Abstract

This paper describes our work on the tasks of opinion detection and opinion polarity detection in blogs. We propose and evaluate a set of topic independent opinion finding evidences for opinion detection task. The approach we adapt is a combination of lexicon-based and machine-learning approaches. In addition, our work includes the classification of topics of TREC Blog collection for years 2006 and 2007. Finally, we analyze the impacts of topic classification on the tasks of opinion detection and opinion polarity detection.

Keywords: Information retrieval, opinion, polarité

