Modèles de langue pour la détection d'opinions dans les blogs

Faiza Belbachir — Mohand Boughanem

Université Toulouse IRIT UMR 5505 CNRS
118 route de Narbonne F-31062 Toulouse cedex 9
*Faiza.Belbachir@irit.fr ** Mohand.Boughanem@irit.fr*

RÉSUMÉ. *Cet article décrit une approche de recherche de documents pertinents vis-à-vis d'une requête et exprimant une opinion. Afin de détecter si un document est porteur d'opinions (i.e; comporte de l'information subjective), nous proposons de le comparer à des sources d'information dont on est sûr qu'elles comportent du contenu de type opinions. L'intuition derrière cela est la suivante, un document ayant une similarité forte avec des sources d'opinions est lui aussi vraisemblablement porteur d'une opinion. Pour mesurer cette similarité nous exploitons des modèles de langues. Nous modélisons le document et la référence porteuse d'opinions par des modèles de langues, nous évaluons ensuite la proximité de ces modèles. Plusieurs expérimentations ont été réalisées sur des collections issues de TREC. Nous proposons de prendre la collection de TREC blog06 comme collection d'analyse et la collection IMDB comme étant la collection de référence.*

ABSTRACT. *This article describes an approach to search relevant documents to the query and expressing an opinion. To detect if a document is opinionated (i.e; contain subjective information), we suggest to compare it with sources of information that contain subjective information. The intuition behind it is the following one, a document having a strong similarity with sources of opinions have an opinion. To measure this similarity we used languages models. We model the document and the reference of opinions using languages models, we estimate then the closeness of these models. Several experiments were realized on collections stemming from TREC. We took the collection of TREC blog06 as collection of analysis and the collection IMDB as being the collection of reference.*

MOTS-CLÉS : *Recherche d'information, blogs, détection d'opinions, modèle de langue*

KEYWORDS: *Information retrieval, blogs, opinions detection, language model*

1. Introduction

Depuis l'avènement d'Internet, de nombreuses formes de contenu ont été générées par les utilisateurs, y compris les pages personnelles, les discussions et les blogs. Certains contenus sont difficiles à copier, rééditer, poster ou redistribuer dans des listes. Ils exigent des permissions spécifiques.

D'où l'explosion des blogs qui sont un moyen plus facile pour l'expression des avis personnels, le partage des sentiments, ou pour commenter sur différents sujets (commercial, politique, etc). A cause de leur popularité les blogs ont attiré beaucoup d'attention dans la communauté du traitement automatique du langage naturel, la recherche d'information et autre (Adar *et al.,* 2005, Agarwal *et al.,* 2008, Ding *et al.,* 2008).

Les approches relatives à cette tache se divisent en deux : celles qui se basent sur le lexique et celles qui se basent sur l'apprentissage. Certains travaux combinent le lexique et l'apprentissage.

Les approches basées sur le lexique utilisent des dictionnaires de mots subjectifs. Si un document contient beaucoup de mots subjectifs alors il est considéré comme un document contenant des opinions (Mishne, 2006, Oard *et al.,* 2006). Les approches basées sur l'apprentissage utilisent différents classifieurs tels que (SVM, Naïve bayes, ect) . Dans un premier temps les classifieurs sont formés par des corpus annotés (opinion ou non opinion) et testés ensuite sur le corpus test (Seki *et al.,* 2007, Zhang *et al.,* 2007).

Ces approches dépendent totalement des dictionnaires ou de la collection d'apprentissage utilisés. Si ces derniers ne sont pas appropriés au langage utilisé dans les blogs alors la détection d'opinions sera faussée. Hors la plupart des travaux utilisent des collections d'apprentissage ou des dictionnaires de mots non représentatifs (formels) du langage utilisé dans les blogs (informel). Pour cela nous proposons dans notre travail d'exploiter des sources d'informations comportant effectivement des informations subjectives de type opinion similaires à celles du langage utilisé dans les blogs. Pour cela, nous définissons un modèle qui permet de représenter une collection porteuse d'opinions.

Afin de déterminer si un document porte une opinion nous mesurons sa similarité avec le modèle des documents de la collection à opinion. Pour ce faire nous proposons de calculer la source d'opinions et les documents sur leurs modèles de langue respectifs.

L'article est constitué de trois sections. Dans la première section nous présentons quelques travaux qui ont adapté leur corpus d'apprentissage à celui des blogs. Dans la deuxième section nous exposons les modèles de langue proposés pour la détection d'opinions. Dans la troisième section nous présentons les collections utilisées et les résultats obtenus, et en dernier nous concluons sur les travaux à venir.

2. Travaux associés

Certains auteurs ont essayé de prendre en considération le fait que le langage utilisé dans les blogs diffère de celui utilisé dans les corpus traditionnels, et ont adapté leurs corpus d'apprentissage à celui des blogs.

Q. Zhang et al. (Zhang *et al.,* 2007) ont utilisé la CME (la Classification en Minimisant l'Erreur) pour assigner un score d'opinions à chaque phrase du blog. Ils ont alors défini quelques caractéristiques basées sur la subjectivité et la pertinence dans toutes les phrases du blog poste. Un classificateur SVM est utilisé pour assigner un score d'opinion à chaque document basé sur les valeurs des caractéristiques définies. Ils ont utilisé un ensemble de phrases subjectives (5000 phrases) et un ensemble de phrases objectives (5000) extraits d'une collection contenant des opinions sur les films, comme données d'apprentissage au classifieur CME. Mais pour le classifieur SVM les données pour l'apprentissage sont ceux de la collection Trec Blog06. Ainsi les blogs sont ordonnés selon leur score final basé sur le score de pertinence multiplié par le score d'opinion.

Amati et al. (Amati *et al.,* 2008) ont proposé de générer automatiquement les mots à opinions des documents pertinents de la collection de Trec Blog06. Dans un premier temps le terme est sélectionné en utilisant « Log Likelihood Ratio » et en prenant en considération sa fréquence dans le document. Ces termes sont ensuite envoyés à un moteur de recherche pour obtenir un score de pertinence. Les documents sont ordonnés selon deux étapes. La première étape consiste à diviser le score d'opinion du document par son score de pertinence et inversement dans la deuxième étape la division se fait entre le score de pertinence du document par son score d'opinion.

W. Zhang et al. (Jia *et al.,* 2008) ont détecté les opinions au niveau des phrases du document. Ils ont utilisé en plus de TREC BLOG06 des ressources à opinions telle que «epinion.com» et ont utilisé un moteur de recherche avec des phrase à opinions comme requêtes. Les premiers documents retournés par ce dernier sont considérés comme des documents comportant des opinions. Ils ont procédé de la même façon pour déterminer les documents qui ne contiennent pas d'opinions, pour cela ils utilisent Wikipédia comme dictionnaire et des requêtes porteuses de mots objectifs (non opinions). Les premiers documents retournés sont considérés comme documents objectifs. Il utilise SVM comme classifieur et ils combinent le score de pertinence avec le score d'opinion..

Yang et al. (Yang *et al.,* 2007) ont aussi utilisé plusieurs ressources telles que : Wilson lexical, ou acronyme modèle, les fréquences des termes dans les documents à opinions, et les termes non fréquents dans les documents ne portant pas d'opinions. Toutes ces caractéristiques sont combinées à travers un classifieur automatique. En dernier une combinaison linéaire entre le score de pertinence et le score d'opinion est faite.

3. Modèles Proposés

Afin de mesurer si un document est porteur d'opinions nous proposons de mesurer sa similarité avec les documents avérés de type opinions. Ceci revient à calculer la probabilité que le document soit généré par le modèle de langue de la collection à opinions. Ce qui revient à calculer une probabilité entre le document et le modèle de référence (opinion). Pour ce faire nous proposons de calculer la probabilité de ressemblance des documents (collection dite d'analyse) avec la collection de référence (collection dite d'opinions) selon deux modèles (dépendant, indépendant). Ensuite nous présentons le modèle de langue de la collection de référence (dite d'opinions). Un score est ensuite calculé selon le modèle (Score_prod_D_R ou Score_KL_D_R) pour réordonner la collection dite d'analyse.

3.1. *Modèle de la collection d'analyse dépendant*

Dans ce modèle nous déterminons un modèle de langue pour la collection dite d'analyse P(w/D) en dépendance avec la collection dite d'opinions (voir équation 1). Une pondération (subjectivité) des termes d'opinions se basant sur le dictionnaire SentiWordNet (Esuli *et al.,* 2006) est calculée (voir équation 2) :

$$P(w|D) = \lambda * P_{ML}(w/D) * subj(w) + (1 - \lambda) * P_{ML}(w/R) * subj(w) \qquad [1]$$

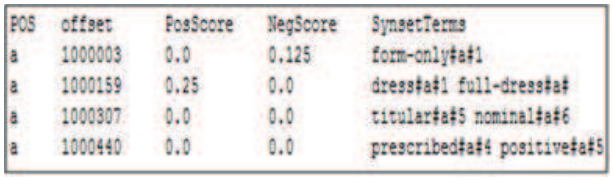
Où $P_{ML}(w|D)$ et $P_{ML}(w|R)$ sont des probabilités qui se basent sur la fréquence du terme du document dans respectivement : le document à analyser, la collection R (dite à opinions) et λ est un paramètre de lissage.

Nous utilisons la ressource lexicale SentiWordNet (SWN) (Esuli *et al.,* 2006) pour pondérer les termes à opinions en d'autres termes, nous calculons la subjectivité des termes dans un document. SWN assigne trois scores (Obj (w), Pos(w), Neg (w)) à chaque synset du WordNet qui représentent respectivement les scores objectif, positif ou négatif. Ces scores sont compris dans l'intervalle [0, 1] et leur somme pour un synset est égale à 1.

Ce processus d'assigner un score à chaque terme permet une détermination sémantique plus précise que celle où on étiquette des termes juste par des étiquettes subjectives ou objectives (pour la tâche d'orientation sémantique) ou bien Fort ou Faible (pour la tâche de force de polarité). La figure 1 montre un exemple de SWN.

Il est aussi très important de noter qu'un terme peut appartenir à plusieurs synsets de SWN et pourrait avoir différentes valeurs de subjectivité dans les différents synsets. Le nombre total de synsets dans lequel un terme apparaît représente le nombre total de sens pour ce terme. Par exemple, le terme brûle a un total de 15 sens.

Un score positif et négatif de 0.0 dans le synset burn#v#12 sunburn#v#1 tandis que dans le synset bite#v#2 burn#v#4 sting#v#1 il porte un score positif de 0.0 et un score négatif de 0.75.



POS	offset	PosScore	NegScore	SynsetTerms
a	1000003	0.0	0.125	form-only#a#1
a	1000159	0.25	0.0	dress#a#1 full-dress#a#
a	1000307	0.0	0.0	titular#a#5 nominal#a#6
a	1000440	0.0	0.0	prescribed#a#4 positive#a#5

Figure 1. *: Exemple de SentiWordNet avec la première colonne : Parts of Speech (POS) («a» signifie adjectif) du Synset, 2ème colonne : Offset du Synset dans WordNet, 3ème Colonne : Score Positif du Synset, 4ème Colonne : Score Négatif du Synset, 5ème Colonne : les Entrées d'un Synset*

Donc en cherchant la subjectivité d'un terme dans SWN, il est plus intéressant d'utiliser la moyenne de subjectivité des termes si nous n'utilisons pas de méthode de désambiguisation tel est notre cas. Nous calculons ainsi la moyenne du score de subjectivité d'un terme en ajoutant le score positif et le score négatif pour tous les sens de ce terme et divisons ensuite le score total par le nombre total des sens du terme (voir l'équation 2). (Missen *et al.,* 2009)

$$Subj(w) = \sum_{s_{i \in sens(w)}} \frac{(Neg(si) + Pos(si))}{|sens(w)|} \quad [2]$$

3.2. *Modèle de la collection d'analyse indépendant*

Dans ce modèle nous déterminons un modèle de langue pour la collection dite d'analyse P(w/D) indépendamment de la collection dite à opinion (R), en se basant sur une simple probabilité des fréquences des termes dans un document (équation 3) :

$$P(w|D) = \frac{frw}{|D|} \quad [3]$$

Où frw est la fréquence d'un terme dans le document et $|D|$ est le nombre de mots dans le document D. Nous déterminons ensuite un modèle de langue pour la collection dite d'opinions (modèle de référence).

3.3. *Modèle de référence*

Le modèle $P(w|R)$ est basé sur la méthode du maximum de vraisemblance, sur la base des fréquences des termes dans la collection R (dite d'opinions).

Un lissage par interpolation (Jelinek *et al.,* 1980) est utilisé pour ne pas obtenir une probabilité égale à zéro (voir équation 4).

$$P(w|R) = \lambda * P_{ML}(w/R) * subj(w) + (1 - \lambda) * P_{ML}(w/C) * subj(w) \quad [4]$$

Où $P_{ML}(w|R)$ et $P_{ML}(w|C)$ sont des probabilités qui se basent sur la fréquence du terme du document dans respectivement : la collection R (dite à opinions) et la collection C (dite d'analyse). λ est un paramètre de lissage, et subj(w) est la pondération de subjectivité définie dans l'équation 2

3.4. *Score de ré-ordonnancement*

Après avoir défini le modèle de langue de la collection dite d'analyse et celui de la collection d'opinion, deux scores de ré-ordonnancement sont calculés. L'un se base sur une mesure de divergence (kl divergence) (Kullback *et al.,* 1951) et qui est représenté dans l'équation 5, et l'autre se basant sur le produit des termes du document (indépendance des termes du document) qui est représenté dans l'équation 6. Le premier score est appliqué pour les documents de la collection dite d'analyse suivant le modèle indépendant, tandis que le deuxième score, il est utilisé pour les documents utilisant le modèle dépendant.

$$score\_KL\_R(D) = \sum_{w \in D} P(w|D) * log \frac{P(w|D)}{P(w|R)} \quad [5]$$

Où $P(w|D)$ et $P(w|R)$ sont les modèles de langue respectivement : du document d'analyse indépendant, de la collection dite à opinion. Plus le score est faible plus le document est similaire à la collection d'opinions.

$$Score\_prod\_R(D) = \prod_{w \in D} P(w|D) \quad [6]$$

Où $P(w/D)$ est le modèle du document dépendant. Plus le score est élevé et plus le document contient des opinions.

4. Évaluation et collections utilisées

4.1. *Collections utilisées*

Nous avons utilisé deux collections l'une de Trec Blog Track (Macdonald *et al.,* 2006) avec 50 topics de 2006, et qui représente la collection à analyser.

La deuxième collection est de IMDB [1] qui est la collection à opinion. Nous calculons ainsi pour chaque blog son modèle de langue dépendant et indépendant des documents de la collection de IMDB.

4.1.1. *Trec Blogs Track*

Cette collection provenant de TREC Blog Track (Macdonald *et al.,* 2006) se compose de plus de 3.2 millions de post blogs extrait durant une période de 11 semaines de Décembre 2005 à Février 2006. TREC propose un ensemble de sujets (50 sujets par an) et un ensemble de jugements de pertinence (qrels) sur un échantillon de blogs tagué : 0 pour définir les blogs non pertinents, 1 pour les blogs pertinents, 2 pour les blogs à opinion négative, 3 pour ceux à opinion mixte et 4 pour ceux à opinion positive.

4.1.2. *IMDB*

Internet Movie Data Base (IMDB) est une base de données en ligne sur le cinéma, la télévision et les jeux vidéo. Toute personne peut poser et partager des avis sur un tel film ou autre. Le site a été créé le 17 Octobre 1990 par Cal Needham et est devenu parmi les sites les plus visités au monde (classé au rang 38 dans le monde) et a plus de 57 millions de visiteurs par mois. L'intérêt de cette collection est qu'elle contient un grand nombre d'opinions, d'avis et de sentiment. Lillian Lee and Bo Pang (Pang *et al.,* 2004) se sont intéressés à cette collection et y ont extrait un ensemble de documents contenant des opinions.

4.2. *Évaluation*

Nous exécutons l'expérimentation en deux phases. Dans la première phase, nous regardons la distribution des scores ($score\_KL\_R$ et $score\_prod\_R$) dans les documents qui contiennent des opinions et dans les documents qui ne contiennent pas des opinions. Dans la deuxième phase, nous exécutons l'évaluation des deux modèles proposés.

4.2.1. *Analyse de caractéristique*

Dans cette partie nous analysons l'efficacité des scores ($score\_KL\_R$ et $score\_prod\_R$) proposés en regardant leurs distributions sur l'ensemble des blogs de TREC (blog à opinions et sur les blogs non opinions). La figure 2 et la figure 3 montrent que les deux scores peuvent être utilisés pour la distinction entre les blogs à opinion ou à non opinions. Car le score ($score\_KL\_R$ ou $score\_prod\_R$) pour les blogs à opinions diverge de celui calculé pour les blogs qui ne contiennent pas d'opinions.

1. `http://www.cs.cornell.edu/people/pabo/movie-review-data/`.

Dans la Figure 2 nous remarquons une différence entre la moyenne des $scores\_KL\_R$ pour les documents à opinions avec la moyenne des $score\_KL\_R$ pour les documents qui ne contiennent pas d'opinions.

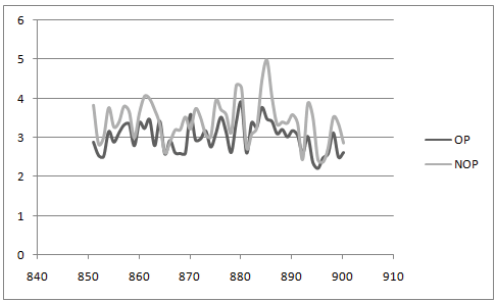


Figure 2. *Le $score\_KL\_R$ dans la collection à opinions (Op) et dans la collection qui ne contient pas d'opinions (Nop)*

Le tableau1 suivant montre un exemple de résultats. Plus le score est faible plus le document est similaire à la collection d'opinions. Nous pouvons dire alors que le $score\_KL\_R$ permet de déterminer les documents d'opinions des documents ne comportant pas d'opinions.

TOPIC	OP	NOP
885	3.487	4.972
869	2.607	3.528
894	2.376	3.524
899	2.513	3.355

Tableau 1. *Les résultats du score $score\_KL\_R$ pour quelque topics de Trec Blog 2006*

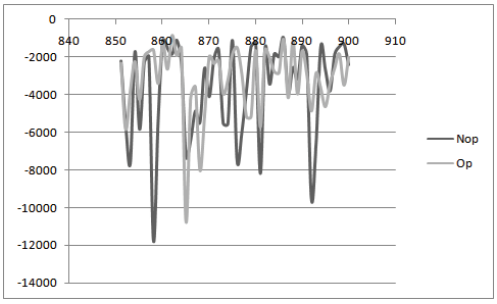


Figure 3. *Le log du $score\_prod\_R$ dans la collection à opinions (Op) et dans la collection qui ne contient pas d'opinions (Nop)*

Dans la figure 3 nous remarquons une différence entre la moyenne des $score\_prod\_R$ pour les documents à opinions avec la moyenne des $scores\_prod\_R$

pour les documents qui ne contiennent pas d'opinions. Plus le score est élevé plus le document est considéré comme un document porteur d'opinions. Mais ceci dit il existe des topics ou la différence est faible comme pour les topics (886, 887, 889). Cela est du à la différence du nombre de documents à opinions par rapport à celui des documents qui ne comportent pas d'opinions.

4.2.2. *Évaluation des deux modèles*

Nous récupérons les 1000 premiers documents pour chaque topic en utilisant le Baseline le plus fort (c'est-à-dire, le Baseline 4 de Trec). Nous réordonnons les documents selon leurs scores finaux $score\_prod\_R$ et $score\_KL\_R$. Nos expérimentations utilisent un facteur de lissage λ égale à 0.6. Les résultats sont donnés dans la Tableau 2.

En analysant ce tableau 2, nous remarquons que la méthode« modèle dépendant » améliore de plus de 3 à 7 % la détection d'opinions par rapport à la méthode basée sur le « modèle indépendant ».

RUN	**OP**	
	MAP	**P@10**
Dépendant	0.1326	0.180
Indépendant	0.1027	0.110

Tableau 2. *: Les résultats de la mesure Map pour l'opinion des deux méthodes dépendant et indépendant.*

Nous pouvons ainsi conclure qu'un blog à opinion peut être généré par un ensemble de documents contenant des opinions. Ceci dit les scores de la mesure MAP et la P@10 (pour les opinions) restent relativement faibles. Ceci peut s'expliquer par le fait que le score de pertinence des documents n'a pas été pris en considération. Le but de notre travail est d'utiliser des modèles de langue pour déterminer les opinions à partir de collection similaire aux blogs telle que (IMDB). Et de pouvoir déduire qu'un blog est généré à partir d'un ensemble de blogs à opinion (méthode dépendance).

5. Conclusion

Nous abordons dans cet article le domaine de la détection d'opinions dans les blogs, nous partons du fait qu'un blog contient des opinions s'il est similaire à un autre blog à opinion. Cette similarité entre les blogs est calculée selon deux modèles de langue nommés « modèle dépendant » et « modèle indépendant ». Nous testons nos deux modèles en considérant la collection de Trec Blog 2006 comme collection d'analyse et la collection de IMDB comme collection à opinions. Nous remarquons que le premier modèle détermine l'opinion plus de 3 à 7% (Map et P@10) que le deuxième modèle.

Afin d'améliorer ces résultats, Il serait intéressant dans un premier temps de pondérer les termes à opinion non pas uniquement par leurs subjectivités mais aussi par d'autres caractéristiques telles que : l'émotivité (nombre d'adjectifs, nombre d'adverbes, nombre de noms et nombre de verbes) (Zhou *et al.,* 2003), l'adressabilité (nombre de : I, my, you, etc...) (Chesley, 2006). Dans un deuxième temps il serait intéressant d'enrichir la collection de référence dite à opinion avec d'autres sources à opinions tel que (epinion.com). Et dans un dernier temps, une combinaison entre le score d'opinion et le score de pertinence (combinaison linéaire, ou avec les rang) (Yue *et al.,* 2007, Montague *et al.,* 2002) sera considérée afin d'aboutir à une meilleur détection d'opinions.

6. Bibliographie

Adar E., Adamic L. A., « Tracking Information Epidemics in Blogspace », *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, IEEE Computer Society, Washington, DC, USA, p. 207-214, 2005.

Agarwal N., Liu H., Tang L., Yu P. S., « Identifying the influential bloggers in a community », *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, ACM, New York, NY, USA, p. 207-218, 2008.

Amati G., Ambrosi E., Bianchi M., Gaibisso C., Gambosi G., « Automatic construction of an opinion-term vocabulary for ad hoc retrieval », *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, Springer-Verlag, Berlin, Heidelberg, p. 89-100, 2008.

Chesley P., « Using verbs and adjectives to automatically classify blog sentiment », *In Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*, p. 27-29, 2006.

Clark S. W. D. H. M., Beresi U. C., « Rgu at the trec blog track », *Text Retrieval Conference*, 2008.

Cronen-Townsend S., Zhou Y., Croft W. B., « Predicting query performance », *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, ACM, New York, NY, USA, p. 299-306, 2002.

Ding X., Liu B., Yu P. S., « A holistic lexicon-based approach to opinion mining », *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, ACM, New York, NY, USA, p. 231-240, 2008.

Ernsting B., Weerkamp W., de Rijke M., « Language Modeling Approaches to Blog Postand Feed Finding », *TREC*, 2007.

Esuli A., Sebastiani F., « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining », *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06*, p. 417-422, 2006.

Hoang L., Lee S.-W., Hong G., Lee J.-Y., Rim H.-C., « A Hybrid Method for Opinion finding Task (KUNLP at TREC 2008 Blog Track) », *TREC*, 2008.

Jelinek F., Mercer R. L., « Interpolated estimation of Markov source parameters from sparse data », *Proceedings of the Workshop on Pattern Recognition in Practice*, 1980.

Jia L., Yu C. T., Zhang W., « UIC at TREC 208 Blog Track », *TREC*, 2008.

Kullback S., Leibler R., « On Information and Sufficiency », *Annals of Mathematical Statistics*, p. $79_8 6, 1951$.

Lafferty J., Zhai C., « Document language models, query models, and risk minimization for information retrieval », *Proceedings of SIGIR*, SIGIR '01, ACM, USA, p. 111-119, 2001.

Liao X., Cao D., Tan S., Liu Y., Ding G., Cheng X., « Combining Language Model with Sentiment Analysis for Opinion Retrieval of Blog-Post », *in* , E. M. Voorhees, , L. P. Buckland (eds), *Proceedings of TREC 2006, Gaithersburg*, vol. Special Publication 500-272, (NIST), 2006.

Macdonald C., Ounis I., « The TREC Blogs06 collection : creating and analysing a blog test collection », 2006.

Manning C. D., Schütze H., *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, USA, 1999.

Mishne G., « Multiple Ranking Strategies for Opinion Retrieval in Blogs - The University of Amsterdam at the 2006 TREC Blog Track », *TREC*, 2006.

Missen M. M. S., Boughanem M., « Using WordNet's Semantic Relations for Opinion Detection in Blogs », *ECIR*, p. 729-733, 2009.

Montague M., Aslam J. A., « Condorcet fusion for improved retrieval », *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, ACM, New York, NY, USA, p. 538-548, 2002.

Oard D. W., Elsayed T., Wang J., Wu Y., Zhang P., Abels E. G., Lin J. J., Soergel D., « TREC 2006 at Maryland : Blog, Enterprise, Legal and QA Tracks », *TREC*, 2006.

Osman D. J., Yearwood J., Vamplew P., « Using Corpus Analysis to Inform Research into Opinion Detection in Blogs », *in* , P. Christen, , P. J. Kennedy, , J. Li, , I. Kolyshkina, , G. J. Williams (eds), *Data Mining and Analytics 2007, Proceedings of the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Queensland, Australia, December 3-4, 2007, Proceedings*, vol. 70 of *CRPIT*, Australian Computer Society, p. 65-75, 2007.

Pang B., Lee L., « A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts », *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, Barcelona, Spain, p. 271-278, July, 2004.

Seki K., Kino Y., Sato S., Uehara K., « TREC 2007 Blog Track Experiments at Kobe University », *in* , E. M. Voorhees, , L. P. Buckland (eds), *Proceedings of TREC 2007, Gaithersburg, Maryland, USA*, vol. Special Publication 500-274, (NIST), 2007.

Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S., « OpinionFinder : a system for subjectivity analysis », *Proceedings of HLT/EMNLP*, HLT-Demo '05, Association for Computational Linguistics, USA, p. 34-35, 2005.

Yang K., Yu N., Zhang H., « WIDIT in TREC 2007 Blog Track : Combining Lexicon-Based Methods to Detect Opinionated Blogs », *TREC*, 2007.

Yue Y., Finley T., Radlinski F., Joachims T., « A support vector method for optimizing average precision », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, ACM, New York, NY, USA, p. 271-278, 2007.

Zhang Q., Wang B., Wu L., Huang X., « FDU at TREC 2007 : Opinion Retrieval of Blog Track », *in* , E. M. Voorhees, , L. P. Buckland (eds), *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, vol. Special Publication 500-274, National Institute of Standards and Technology (NIST), 2007.

Zhou L., Twitchell D. P., Qin T., Burgoon J. K., Nunamaker Jr. J. F., « An Exploratory Study into Deception Detection in Text-Based Computer-Mediated Communication », *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track1 - Volume 1*, HICSS '03, IEEE Computer Society, Washington, DC, USA, p. 44.2-, 2003.