L'apport des Entités Nommées pour la classification des opinions minoritaires

Amel Fraisse, Patrick Paroubek, Gil Francopoulo

L'apport des Entités Nommées pour la classification des opinions minoritaires

Amel Fraisse[1] Patrick Paroubek[1] Gil Francopoulo[2]
(1) LIMSI-CNRS, Bât. 508 Université Paris-Sud, 91403 Orsay Cedex, France
(2) TAGMATICA, 126 rue de Picpus, 75012 Paris France
fraisse@limsi.fr, pap@limsi.fr, gil.francopoulo@tagmatica.com

RÉSUMÉ

La majeure partie des travaux en fouille d'opinion et en analyse de sentiment concerne le classement des opinions majoritaires. Les méthodes d'apprentissage supervisé à base de n-grammes sont souvent employées. Elles ont l'inconvénient d'avoir un biais en faveur des opinions majoritaires si on les utilise de manière classique. En fait la présence d'un terme particulier, fortement associé à la cible de l'opinion dans un document peut parfois suffire à faire basculer le classement de ce document dans la classe de ceux qui expriment une opinion majoritaire sur la cible. C'est un phénomène positif pour l'exactitude globale du classifieur, mais les documents exprimant des opinions minoritaires sont souvent mal classés. Ce point est un problème dans le cas où l'on s'intéresse à la détection des signaux faibles (détection de rumeur) ou pour l'anticipation de renversement de tendance. Nous proposons dans cet article d'améliorer la classification des opinions minoritaires en prenant en compte les Entités Nommées dans le calcul de pondération destiné à corriger le biais en faveur des opinions majoritaires.

ABSTRACT

Improving Minor Opinion Polarity Classification with Named Entity Analysis

The main part of the work on opinion mining and sentiment analysis concerns polarity classification of majority opinions. Supervised machine learning with n-gram features is a common approach to polarity classification, which is often biased towards the majority of opinions about a given opinion target, when using this kind of approach with traditional settings. The presence of a specific term, strongly associated to the opinion target in a document, is often enough to tip the classifier decision toward the majority opinion class. This is actually a good thing for overall accuracy. Howeverm documents about the opinion taget, but expressing a polarity different from the majority one, get misclassified. It is a problem if we want to detect weak signals (rumor detection) or for anticipating opinion reversal trends. We propose in this paper to improve minor reviews polarity classification by taking into account Named Entity information in the computation of specific weighting scheme used for correcting the bias toward majority opinions.

MOTS-CLÉS : Fouille d'opinions, Opinion minoritaires, Entités Nommées, Apprentissage, N-grammes, Pondération.

KEYWORDS: Opinion Mining, Minor Opinion, Named Entities, Machine Learning, N-grams, Weighting Scheme.

1 Introduction

Il est devenu de nos jours très facile d'assembler de grandes quantités de textes d'opinion à partir des réseaux sociaux, de forums et de sites de critique en ligne pour construire un classifieur de documents basé sur les opinions, qui fonctionnera avec un niveau de performance suffisant pour une utilisation industrielle. Cependant, un tel système est souvent biaisé en faveur des opinions majoritaires exprimés à propos d'une cible particulière présente dans les données d'entrainement. Si nous utilisons un tel système pour analyser de nouveaux documents concernant la même cible, il est très vraisemblable qu'ils seront affectés par le classifieur au courant d'opinion majoritaire, essentiellement du fait de la présence de termes spécifiques à la cible de l'opinion dans ces documents. Bien sûr cela s'applique à n'importe quel type de document, en particulier les critiques de produits, films etc. Par exemple, si l'on cherche à classer un document qui parle d'un film à succès, la simple mention du titre, d'un acteur de la distribution, du producteur, ou du metteur en scène, sera suffisante pour que le classifieur lui assigne une catégorie positive.

Paradoxallement, ce biais en faveur de l'opinion majoritaire favorise l'exactitude globale des systèmes d'analyse d'opinion lorsque seules deux ou trois classes d'opinion sont considérées, car la distribution des différents types de documents des données d'entraînement est supposée refléter la distribution présente des différents types de document des données de test. En fait, c'est une considération qui a même servi d'hypothèse de travail pour constituer le corpus d'apprentissage. Si une cible d'opinion est majoritairement positive dans les données d'entraînement on s'attend a ce que les données de test contiennent plus de documents à teneur positive que de documents négatifs à propos de cette cible.

Mais dans cetains cas, il est souhaitable d'avoir un système qui soit aussi capable de déterminer correctement la polarité d'un document exprimant une opinion minoritaire, par exemple lorsque l'on cherche à détecter des signaux faibles (pour la détection de rumeur) ou bien si l'on cherche à anticiper des retournements d'opinion majoritaire. Ce dernier point est particulièrement stratégique pour toutes les industries reposant sur la fourniture de service, où l'on cherche à fidéliser ses abonnés. Un système de fouille d'opinion capable d'effectuer un classement correct des opinions minoritaires peut être comparé à un expert qui prend des décisions de classement uniquement en fonction des opinions exprimées dans un document particulier à propos d'une cible, sans tenir compte de l'opinion générale sur cette cible.

2 Schémas de pondération

Nous représentons un document donné d comme un ensemble de traits : $d = \{g_1, g_2, ..., g_k\}$, nous définissons son vecteur de poids $w_d = \{w(g_1), w(g_2), ..., w(g_k)\}$, où $w(g_i)$ est le poids du trait g_i dans le document d. Dans un premier temps, nous utilisons les deux schémas de pondération les plus utilisés dans le domaine de l'analyse de sentiment : Binaire et DELTA-TFIDF (Martineau et Finin, 2009) (Paltoglou et Thelwall, 2010). Ensuite, nous utilisons les trois schémas de pondérations proposés par (Pak et Parboubek, 2011) pour améliorer la classification des opinions minoritaires. Le principe de base de ces trois métriques consiste à réduire l'importance des traits qui pourraient introduire un biais dans le classement d'une critique minoritaire. Comme décrit dans (Pak, 2012), la première métrique est basée sur la fréquence moyenne d'un trait. La deuxième métrique appelée *proportion d'entité (ep)* est basée sur les occurrences d'un trait à

travers l'ensemble des entités e [1], comparativement à sa fréquence d'apparition dans l'ensemble de documents. La troisième métrique, combine la fréquence moyenne d'un trait et sa proportion d'entité.

2.1 Fréquence moyenne d'un trait

La fréquence moyenne d'un trait (agv.tf) est le nombre moyen de fois qu'un trait apparaît dans un document, $\{d|g_i \in d\}$ est l'ensemble de documents qui contient g_i, (Pak, 2012).

$$\text{avg.tf}(g_i) = \frac{\sum_{\{d|g_i \in d\}} \text{tf}(g_i)}{\|\{d|g_i \in d\}\|} \tag{1}$$

La normalisation à base de fréquence moyenne d'un trait est basée sur l'observation que les auteurs de critiques ont tendance à utiliser un vocabulaire riche quand ils expriment leur attitude par rapport à un film ou un produit. Ainsi, les traits exprimant des sentiments comme remarquable (outstanding) ou adorable (lovingly) ont une fréquence moyenne proche ou égale a 1, tandis que les traits non subjectifs ont une fréquence moyenne plus élevée. Afin de normaliser le vecteur représentatif d'un document qui associe à chaque trait présent dans le document un poids représentatif de son importance, nous divisons chaque poids par la fréquence moyenne du trait correspondant (Pak, 2012) :

$$w(g_i)^* = \frac{w(g_i)}{\text{avg.tf}(g_i)} \tag{2}$$

2.2 Proportion d'entité

La proportion d'entité (ep) est la proportion des occurrences d'un trait par rapport aux différentes entités comparativement à la fréquence des documents (Pak, 2012).

$$\text{ep}(g_i) = \log\left(\frac{\|\{e|g_i \in e\}\|}{\|\{d|g_i \in d\}\|} \cdot \frac{\|D\|}{\|E\|}\right) \tag{3}$$

où $\{e|g_i \in e\}$ est l'ensemble des entités qui contiennent g_i, $\|D\|$ est le nombre total de documents, $\|E\|$ est le nombre total d'entité. La normalisation de proportion d'entité favorise les traits qui apparaissent dans nombreuses entités mais rarement dans l'ensemble de documents. Nous distinguons trois types de traits : (a) le vocabulaire d'une e, tels que le numéro de série d'un film, la puissance d'une machine [2]. Ce type de trait est associé à peu d'entité et donc devraient apparaître dans peu de documents. La valeur de ep devrait être proche de celle de la constante de normalisation $NC = \frac{\|D\|}{\|E\|}$, (b) les mots-outils, tels que les déterminants et les prépositions, devraient apparaître dans presque tous les documents, et donc associés à presque toutes les entités. La valeur de ep sera proche de celle de la NC et enfin (c) les termes subjectifs, tels que « remarquable » ou « adorable », qui devraient apparaître associés à beaucoup de produits et dans un nombre relativement restreint de documents, car les auteurs utilisent un vocabulaire varié. La valeur de ep sera plus grande que la constante de normalisation NC. Pour normaliser

1. (Pak, 2012) désigne par *entité (e)* l'ensemble de documents ayant la même cible d'opinion.
2. Les termes du vocabulaire d'entité ne sont pas reconnus, en général, comme entités nommées.

le vecteur représentatif d'un document, nous multiplions chaque poids associé à un trait par sa proportion d'entité (Pak, 2012).

$$w(g_i)^* = w(g_i) \cdot \text{ep}(g_i) \tag{4}$$

Le troisième schéma de pondération proposé par (Pak, 2012), consiste à combiner la fréquence moyenne d'un trait et la proportion d'entité selon la formule suivante :

$$w(g_i)^* = w(g_i) \cdot \frac{\text{ep}(g_i)}{\text{avg.tf}(g_i)} \tag{5}$$

2.3 Notre contribution : pondération des entités nommées

C'est en effet en faveur du développement de la tâche d'extraction d'information que la tâche de reconnaissance des entités nommées (EN) est apparue. Cette tâche a gagné en maturité et s'est précisée grâce à la série des conférences MUC (Message Understanding Conferences) (Grouin *et al.*, 2011). Ensuite, le concept des entités nommées a été repris dans le cadre des campagnes d'évaluation du projet européen QUAERO (Galibert *et al.*, 2012).

La majorité des modèles d'opinion comprennent 3 éléments : l'expression d'opinion, la source, et la cible de l'opinion (Paroubek *et al.*, 2010). Nous nous intéressons ici à la cible, à laquelle, dans la plupart des cas, on fait référence au moyen d'entités nommées (personne, produit, organisation, lieu etc.) et auxquelles est souvent associé un ensemble d'entités nommées contextuelles, propre à la cible, comme par exemple la distribution d'un film ou le nom de son metteur en scène.

Les systèmes de fouille d'opinion et plus particulièrement de classement en polarité, basés sur les approches traditionnelles, c'est-à-dire les approches à base d'apprentissage automatique supervisé utilisant les simples modèles à sac-de-mots (Pak, 2012), ont tendance à s'appuyer sur les traits spécifiques des entités et, par voie de conséquences, ils sont biaisés en faveur des opinions majoritaires présents dans les données d'apprentissage. En particulier les traits réprésentant des EN utilisés pour référencer la cible de l'opinion sont identifiés par le système comme indicateurs de polarité pour les opinions majoritaires. Prenons l'exemple d'un film qui a eu un grand succès comme *AVATAR*, non seulement la mention du titre du film, dans le commentaire, qui pourrait entraîner une classification positive du commentaire mais aussi la citation du nom du directeur du film *James Cameron*, et cela même dans le cas, où il s'agit d'un commentaire négatif sur le film. En outre, les entités nommées ne font pas parti du vocabulaire général pour l'expression d'opinion et de sentiments. De notre point de vue, un système de fouille d'opinion doit être capable de faire la distinction entre les indicateurs d'opinion exprimés de façon explicite, dans l'expression de l'opinion, et ceux qui sont liés à la présence des traits contextuels [3] comme par exemple les EN. Afin d'améliorer la classification des opinions minoritaires, il faut donc réduire l'importance des traits contextuels qui souvent, introduisent un biais dans le processus de classification de ce type d'opinion. Nous proposons donc de compléter la normalisation des schémas de pondération de (Pak, 2012) en se basant sur un système de reconnaissance des EN. Dans un premier ensemble d'expérimentation, nous avons procédé de la façon suivante : si un trait g_i est reconnu comme une EN alors son poids est écarté du vecteur de poids du document (voir Eq. 6).

3. Ensemble de traits associé à une cible d'opinion et qui définissent le contexte de la cible. Par exemple, les traits *Quentin Tarantino* et *Jamie Foxx* apparaissent souvent dans le contexte d'une critique du film *Django*.

$$w_d^{NE} = w_d \ \setminus \{w(g_i), \quad g_i \in NE\} \tag{6}$$

3 Expérimentations

3.1 Données

Nous utilisons le jeu de données : Large Movie Review Dataset (Maas *et al.*, 2011) qui a été utilisé par le passé pour des recherches en analyse de sentiment. Il contient 50 000 critiques de film répartis selon une proportion égale entre les critiques négatives et les critiques positives. Pour préparer le jeu de données, nous avons suivi la procédure décrite dans (Pak, 2012). Pour chaque film, nous avons pris trois documents pour le test et sept pour l'apprentissage. Ces valeurs ont été choisies de manière heuristique afin de maximiser le nombre total de critiques. La constitution des données est illustrée dans la figure 1. Pour séparer l'ensemble de données en sous-ensembles

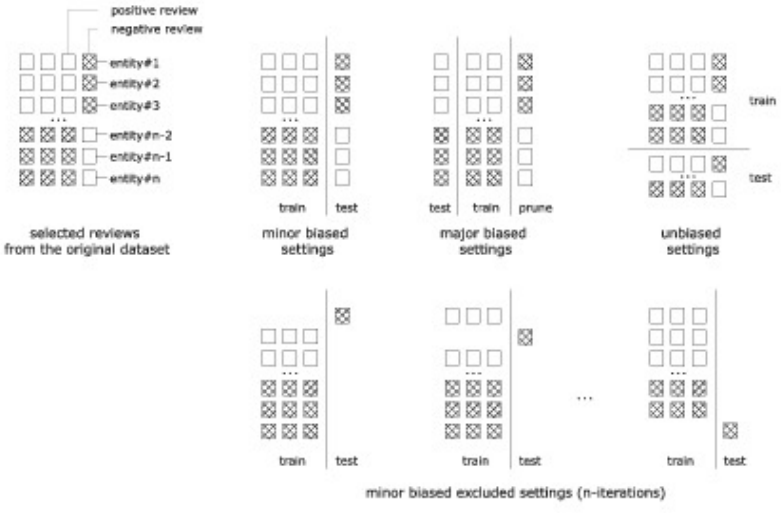

FIGURE 1 – Processus de composition de jeu de données (Pak, 2012).

d'apprentissage et test, au départ, nous avons groupé toutes les critiques par *e* (c'est-à-dire le film), identifiée par un identifiant unique dans l'ensemble des données. A partir de ces groupes, pour chaque *e* nous avons choisi toutes les critiques d'une polarité dominante dans ce groupe et nous les avons transféré dans le corpus d'apprentissage. Les critiques restantes de chaque groupe sont transférées dans le corpus de test. Nous appelons ce corpus "biaisé de manière minoritaire" car le corpus de test contient des critiques avec une polaritée minoritaire. Afin de prouver que la baisse de performance est due effectivement aux traits biaisés, nous avons construit un corpus de test composé des mêmes critiques mais ré-organisé de telle manière que les critiques aient la même polarité que la polarité dominante dans le corpus d'apprentissage pour chaque *e*. Nous appelons ce corpus "biaisé de manière majoritaire". Enfin, nous avons construit

le corpus "non-biaisé" de telle manière que le corpus du test et le corpus d'apprentissage ne contiennent aucun document en commun.

3.2 Résultats

Pour nos expérimentations, nous avons utilisé la bibliothèque LIBLINEAR avec un noyau linéaire et un paramétrage par défaut (Fan *et al.*, 2008). Les entités nommées ont été marquées par G. Francopoulo avec l'outil industriel TagParser de TAGMATICA (Francopoulo et Demay, 2011). En premier lieu, nous prouvons l'effet négatif des traits contextuels et spécifiques aux *e* sur l'exactitude de classification des critiques minoritaires. Nous avons effectué des expérimentations sur trois variantes des corpus : non-biaisé (unb), biaisé de manière minoritaire (minb), biaisé de manière majoritaire (majb). Nous avons utilisé les unigrammes (uni) et les bigrammes (bi) avec des poids : binaire (bin) et Delta-TFIDF. Les résultats sur l'exactitude de la classification selon les corpus et les n-grammes sont présentés dans la table 1.

	unb.	Δ	minb.	Δ	majb.	Δ
Unigrammes + binaire						
bin	80.7		69.4		83.4	
avg.tf	81.5	+0.8	72.3	+2.9	84.8	+1.4
ep	80.1	-0.6	71.3	+1.9	83.5	+0.1
comb	80.7	+0.0	73.0	+3.6	84.4	+1.0
comb.ex.NE	79.5	-1.2	73.6	+4.2	84.6	+1.2
Unigrammes + Delta TF-IDF						
Delta TF-IDF	83.3		63.5		89.2	
avg.tf	81.1	-2.2	69.4	+5.9	87.6	-1.6
ep	82.3	-1.0	67.2	+3.7	87.8	-1.4
comb	81.7	-1.6	69.0	+5.5	87.5	-1.7
comb.ex.NE	81.2	-2.1	**71.4**	**+7.9**	87.5	-2.1
Bigrammes + binaire						
bin	79.6		71.9		83.5	
avg.tf	79.7	+0.1	72.8	+0.9	84.0	+0.5
ep	80.3	+0.7	74.0	+2.1	84.2	+0.7
comb	80.8	+1.2	74.9	+3.0	84.6	+1.1
comb.ex.NE	81.1	+1.5	76.1	+4.2	84.8	+1.3
Bigrammes + Delta TF-IDF						
Delta TF-IDF	83.0		69.9		87.6	
avg.tf	82.9	-0.1	76.0	+6.0	86.1	-1.5
ep	83.2	+0.2	74.4	+4.5	86.2	-1.4
comb	83.3	+0.3	75.1	+5.2	85.8	-1.8
comb.ex.NE	83.9	+0.9	**78.1**	**+8.2**	85.2	-2.4

TABLE 1 – Exactitude de classification des critiques des films en utilisant les différents schémas de normalisation.

Impact des traits contextuels et spécifiques aux entités. En observant la table 1, nous constatons que les traits associés aux entités provoquent une baisse des performances sur le corpus biaisé de manière minoritaire quand on le compare avec le corpus non-biaisé (unb vs. minb). Nous observons aussi une augmentation des performance sur le corpus biaisé de manière majoritaire malgré une taille du corpus d'apprentissage plus petite (unb vs majb). Cela montre que notre classifieur apprend à associer les traits contextuels et spécifiques à une *e* à sa polarité majoritaire. Les résultats sont similaires d'un corpus à l'autre, d'une variante à l'autre et d'une propriété à l'autre. Le Delta TFIDF bien qu'améliorant l'exactitude globale provoque des mauvaises classifications de critiques minoritaires car il donne de l'importance aux traits spécifiques y compris donc les traits représentants des EN. Nous l'observons en comparant les résultats en utilisant Delta TFIDF (uni + Δ et bi + Δ) sur le corpus biaisé de manière minoritaire avec les corpus non-biaisés et biaisés de manière majoritaire. Enfin, nous avons évalué les effets du schéma de normalisation proposé sur l'exactitude de la classification. Ainsi que nous l'avons observé sur les précédentes expérimentations, le fait d'exclure les poids des traits représentant des EN augmente la performance comme présenté dans les parties mises en exergue dans la table 1.

4 Conclusion

Les méthodes que nous avons proposées dans cet article permettent de diminuer l'importance des EN spécifiques à une cible d'opinion particulière, en normalisant leur poids dans le vecteur de poids qui est utilisé par les représentations classiques à base de n-grammes en apprentissage automatique. Les évaluations que nous avons effectuées sur des jeux de donnés spécialement construit à partir de jeux de données standard pour tester nos hypothèses, ont montré que le classement des documents exprimant des opinions minoritaires est grandement amélioré (+8%), ce qui prouve que notre mode de pondération prenant en compte les EN a un impact positif sur la mesure d'exactitude de classification pour les documents d'opinion minoritaires, une nécessité pour la détection de signaux faibles ou l'anticipation de renversement de tendance. Il faut cependant noter que l'accroissement de performance n'est pas aussi important pour les modèles à base de bigrammes (+1%) entraînés avec des données naturellement biaisées, mais il reste positif, ce qui prouve que notre mode de pondération fonctionne au moins aussi bien que les approches classiques sur ces données.

Références

BURGER, J., PALMER, D. et HIRSCHMAN, L. (1998). Named entity scoring for speech input. *In Proceedings of the 17th ICCL*, pages 201–205, Montréal, Québec, Canada. ACL. http ://dx.doi.org/10.3115/980845.980878.

FRANCOPOULO, G. et DEMAY, F. (2011). A deep ontology for named entities. *In Proceedings of the Int. Conf. on Computational Semantics, Interoperable Semantic Annotation workshop*. ACL. http ://tagmatica.fr/publications/FrancopouloACLISOWorkshopWithinInternationalConferenceOnComputational

GALIBERT, O., ROSSET, S., GROUIN, C., ZWEIGENBAUM, P. et QUINTARD, L. (2012). Extended named entity annotation on ocred documents : From corpus constitution to evaluation campaign. *In Proceedings of the 8th LREC*, pages 3126–3131, Istanbul, Turkey. ELDA.

GROUIN, C., S., S. R., ZWEIGENBAUM, P., FORT, K., GALIBERT, O. et QUINTARD, L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Proceedings of the Fifth Law Workshop (LAW V)*, pages 92–100, Portland, Oregon. ACL.

MARTINEAU, J. et FININ, T. (2009). Delta tfidf : An improved feature space for sentiment analysis. *In Proceedings of the Third AAAI Int. Conf. on Weblogs and Social Media*, San Jose, CA. AAAI Press.

MAAS, A.L. DALY, R.E. PHAM, P.T HUANG, D., NG , A.Y et POTTS, C.(2011). Learning word vectors for sentiment analysis. *In Proceedings of the 49th Annual Meeting of the ACL*, pages 142–150, Portland, Oregon, USA. ACL. http ://www.aclweb.org/anthology/P11-1015.

FAN, R.-E CHANG, K.-W HSIEH, C.-J WANG, X.-R et LIN, C.-J (2008). Liblinear : A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874. URL http ://portal.acm.org/citation.cfm ?id=1390681.1442794.

PAK, A. (2012). *Automatic, Adaptive,and Applicative Sentiment Analysis*. Thèse de doctorat, Thèse de l'École Doctorale d'Informatique de l'Université Paris-Sud, Orsay.

PAK, A. et PARBOUBEK, P. (2011). Normalization of term weighting scheme for sentiment analysis. *In Language and Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 415–419, Poznań, Poland.

PALTOGLOU, G. et THELWALL, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. *In Proceedings of the 48th Annual Meeting of the ACL*, pages 1386–1395, Morristown, NJ, USA,. ACL.

PAROUBEK, P., PAK, A. et MOSTEFA, D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valetta, Malta. ELDA.