Combiner lexique et régression logistique dans la classification d'avis laissés sur le Net : une étude de cas

Stefania Pecore, Jeanne Villaneau, Farida Saïd

Combiner lexique et régression logistique dans la classification d'avis laissés sur le Net : une étude de cas

Stefania Pecore[1] Jeanne Villaneau[1] Farida Saïd[2] Pierre-François Marteau[1]
(1) IRISA, UMR 6074, Université de Bretagne Sud, 56000, France
(2) LMBA, UMR 6205, Université de Bretagne Sud, 56000, France
stefania.pecore (jeanne.villaneau, farida.said,
pierre-francois.marteau)@univ-ubs.fr

RÉSUMÉ

L'article compare deux approches en sac de mots pour classifier des avis laissés sur des sites internet de langue française. La première, sans lexique et uniquement statistique, repose sur la régression logistique. La seconde repose sur un lexique d'opinion qui réunit les mots issus de la régression logistique avec une liste de noms, adjectifs et verbes courants annotés manuellement. Les résultats montrent l'intérêt que présente la régression logistique pour enrichir un lexique d'opinion. Par ailleurs, l'analyse des résultats permet de conjecturer les pistes à privilégier pour pallier les insuffisances des approches en sac de mots, particulièrement l'étude de la négation.

ABSTRACT

Combining lexicon and logistic regression in opinion mining : a case study

We compare two bag-of-words approaches aimed to classify reviews extracted from French websites. Our first approach is exclusively statistical : it combines bag-of-words techniques with logistic regression, regardless of any lexicon. Contrastingly, our second approach makes use of an opinion lexicon consisting of a list of manually annotated words (nouns, adjectives, verbs), together with a list of words selected from a logistic regression. Our results show logistic regression as a good technique to improve and enrich opinion lexicons. Furthermore, our analysis suggests directions to follow in order to address the shortcomings of the bag-of-words approach. In particular, it emphasizes the importance of taking into account the contribution of negation in opinion mining.

MOTS-CLÉS : analyse des opinions, analyse des sentiments, lexique d'opinion.

KEYWORDS: opinion analysis, sentiment analysis, sentiment lexicon.

1 Introduction et état de l'art

Étant donnée l'importance des enjeux concernés, les recherches en détection d'opinion se sont multipliées depuis le milieu des années 90 et le sujet a été désormais suffisamment exploré pour que l'on puisse en saisir à la fois la complexité et la diversité. Un signe de cette maturité est donné par les articles et chapitres qui proposent une synthèse du domaine et de ses différents aspects (Westerski, 2007; Pang & Lee, 2008; Feldman, 2013; Breck & Cardie, 2016). Tous les auteurs soulignent l'importance du lexique et de la difficulté posée par son acquisition : Feldman (2013) écrit qu'il est *the most crucial resource* et souligne que chaque sous-domaine spécifique correspond à son propre lexique. Brek & Cardie précise que si certains termes sont explicitement péjoratifs ou mélioratifs,

d'autres dépendent fortement du contexte dans lequel ils apparaissent. Plusieurs méthodes semi-automatiques d'acquisition d'un lexique d'opinion ont été proposées, le plus souvent par extension d'un ensemble de mots déjà annotés. La méthode peut être statistique et s'appuyer sur les contextes d'utilisation (Vincze & Bestgen, 2011; Qiu *et al.*, 2011) ou reposer sur les relations de synonymie ou d'antonymie (Kamps *et al.*, 2004).

Les travaux présentés concernent la classification automatique d'avis laissés sur des sites francophones. L'un des buts est de voir comment certaines approches statistiques – en l'occurrence les variables rendues par la régression logistique – permettent de trouver des lemmes représentatifs de l'opinion dans un corpus donné et si ces lemmes peuvent se combiner avec l'emploi d'un lexique d'opinion généraliste. Les expérimentations présentées section 2 réunissent deux approches en sac de mots. La première, statistique, utilise la régression logistique ; la seconde associe les mots utilisés par la première avec des lexiques d'opinion généraux composés d'adjectifs, de noms et de verbes fréquents. Deux résultats essentiels s'en dégagent.

– Dans la seconde approche, le lexique rendu par la régression logistique est plus efficace que les lexiques généraux mais cependant, sa complétion permet une amélioration des résultats.
– L'étude des mots sélectionnés par la régression logistique et des erreurs du classifieur donne des pistes pour dépasser l'approche en sacs de mots. Elles sont décrites section 3.

2 Expérimentations et résultats

Le corpus utilisé a été collecté par Vincent & Winterstein (2013). Il est composé d'avis déposés sur trois sites différents : *TripAdvisor.fr*, *Allocine.fr*, *Amazon.fr*, et ses données se refèrent respectivement à trois types de produits : hôtels, films et livres. Les critiques sont accompagnées par des notes qui vont de 1 à 5. Les créateurs du corpus ont choisi de sélectionner les avis qui correspondent aux notes extrêmes, c'est-à-dire 1 (très négatifs) et 5 (très positifs) dans les proportions 1000/1000 pour les hôtels, 800/800 pour les films et 576/858 pour les livres. Pour plus de détails, on pourra se reporter à l'article des auteurs (Vincent & Winterstein, 2013).

L'approche de Vincent et Winterstein est une approche en sac de mots par apprentissage automatique qui ne nécessite pas de lexique d'opinion. Nous avons reproduit leurs expérimentations avec la régression logistique (package *glmnet* de R) et obtenu les résultats indiqués dans la première colonne du tableau 1, un peu inférieurs à ceux rapportés par les auteurs, probablement parce que nous n'avons réalisé qu'un pré-traitement très sommaire du corpus. Les trois premiers quarts du corpus (dans chaque thème et pour chaque score) ont été utilisés pour l'entraînement et le dernier quart comme pour les tests. La petite taille du corpus nous a conduit à tester une validation croisée qui n'a pas modifié les résultats de façon sensible. Les expérimentations suivantes ont été menées sur le même corpus de tests. On peut remarquer la très forte différence entre les résultats obtenus sur le corpus des hôtels avec ceux qui correspondent aux avis sur les films et livres.

Notre objectif était de comparer cette approche en sac de mots statistique et sans lexique avec une approche en sac de mots basée sur un lexique d'opinion, par une simple sommation des scores des lemmes qui composent le texte, suivie d'une normalisation en divisant le résultat par le nombre de lemmes porteurs d'opinion. La détermination des scores des mots en opinion a utilisé trois ressources.

1. La première ressource (SB par la suite) est un lexique annoté en émotion de 3082 lemmes obtenu par la réunion de deux lexiques annotés en émotion.

– Les normes émotionnelles *ValEmo* d'Arielle Syssau et Noëlle Font (Syssau & Font, 2005) proviennent d'une étude psycho-linguistique destinée à déterminer l'émotion associée aux mots. Ainsi, 604 lemmes issus de deux normes d'association verbale en langue française (Ferrand & Alario, 1998; Ferrand, 2001) ont été évalués par 600 personnes sur une échelle de 11 points (*-5/+5*).
– Les normes de Nadja Vincze et Yves Bestgen (Vincze & Bestgen, 2011) sont le produit de l'extension de normes déjà existantes en langue française. La norme *F-POL* est composée de 3252 mots évalués par une trentaine de personnes sur une échelle à 7 points, de 1 (très désagréable) à 7 (très agréable).

L'indice de corrélation de Pearson entre les scores des lemmes communs aux deux normes étant égal à 0,943, nous avons réuni les deux ressources en ramenant les scores du lexique de Vincze et Bestgen sur l'échelle *-5/+5*.

2. La deuxième ressource lexicale (RLex) est la liste des variables (en l'occurrence les lemmes) avec les coefficients correspondants, sélectionnées par la régression logistique (paquet *glmnet* de R) pour réaliser la classification. Nous avons normalisé les coefficients sur l'échelle *-5/+5*.
3. La troisième ressource (Lex) est une liste d'environ 1500 noms, 950 adjectifs et 400 verbes très courants, choisis pour leur caractère généralement péjoratif ou mélioratif, que nous avons annotés manuellement sur l'échelle *-5/5*.

Les résultats présentés dans les colonnes centrales (2 à 6) du tableau 1 sont le résultat des classifications obtenues en utilisant une approche en sac de mots et les trois ressources lexicales précédentes : SB, RLex et Lex. Plus précisément, on y trouve la F-mesure obtenue pour les trois corpus relatifs aux hôtels, films et livres, en fonction du lexique de mots utilisé. La première ligne du tableau précise le nombre de mots du lexique correspondant. La dernière colonne du tableau correspond aux derniers résultats d'une approche basée sur la troisième ressource lexicale (Lex) et une analyse de surface du texte (chunking). Des précisions sur ce travail, actuellement en cours, sont données dans la section 3.

	Rég. Log.	SB	Lex	RLex	RLex+SB	RLex+Lex	Ch+Lex
Méthode	Stat	SdM : sommation des scores des lemmes					Chunking
Nb. Mots	0	3082	2894	283	3197	3054	2894
Hôtels	0,940	0,826	0,871	0,930	0,905	**0,955**	0,892
Films	0,856	0,638	0,642	0,713	0,686	**0,741**	0,726
Livres	0,804	0,638	0,677	0,780	0,792	**0,795**	**0,808**

TABLE 1 – Résultats (F-mesure) obtenus pour chacune des différentes méthodes et pour chacun des trois corpus.

– D'une manière générale, les scores obtenus avec une approche en sac de mots par simple sommation des valeurs en opinion des lemmes sont au mieux comparables à ceux de la régression logistique.
– Par ailleurs, dans chacun des trois corpus, le vocabulaire des presque 3000 adjectifs, verbes et noms (*Lex*) se révèle beaucoup moins efficace que le lexique de 283 mots élaboré à partir des termes utilisés par la régression logistique (*RLex*).
– L'utilisation du lexique émotionnel obtenu par combinaisons des études de Syssau et de Bestgen donne des résultats décevants. Une analyse des textes avec un score erroné fait apparaître la différence qui sépare lexique émotionnel et lexique d'opinion. Par exemple, dans la mesure où ils évoquent en général le repos et le calme, les mots *dormir* et *chambre* ont été notés très positivement

dans le lexique affectif (respectivement 4,02 et 2,35). Or, le mot *chambre* n'a pas de connotation positive dans le corpus des hôtels et *dormir* est plutôt connoté négativement dans celui des films...
– Dans les trois corpus, les lexiques les plus efficaces sont ceux qui combinent les mots utilisés par la régression logistique avec le lexique des adjectifs et des noms annotés en opinion, une étude plus fine ayant montré que les verbes sont au mieux inutiles.
– Quelle que soit la méthode choisie, les résultats sont fortement dépendants du corpus. Les meilleurs s'observent dans le corpus des hôtels.
– Bien qu'elle soit encore loin d'être aboutie, l'approche qui utilise le chunking obtient d'ores et déjà les meilleurs scores sur le corpus qui correspond aux avis donnés sur les livres.

3 Discussion

La régression logistique par une approche de type *sac de mots* recherche la meilleure combinaison de termes susceptible de faire la séparation entre les avis positifs et négatifs. En ce sens, les résultats de ces expérimentations lexicales sont cohérents. Par ailleurs, l'étude des résultats donne des indications pour dépasser les limites de ce type d'approche.
– Les mots *ne*, *pas*, *plus* sont utilisés par la régression logistique avec des coefficients fortement négatifs, ce qui montre que les négations sont davantage présentes dans les avis négatifs que dans les positifs et confirme leur importance. Le sujet a donné lieu à bon nombre d'études spécifiques, essentiellement consacrées à la langue anglaise (Jia *et al.*, 2009; Wiegand *et al.*, 2010; Dadvar *et al.*, 2011; Hogenboom *et al.*, 2011; Zhang *et al.*, 2012). La détection de la négation et de sa portée est en effet un problème complexe (Wilson *et al.*, 2005) ; en langue française par exemple, l'indice le plus fiable de la négation est le *ne* mais il est souvent absent dans la langue courante.
– Certains mots sont utilisés par la régression logistique avec des scores qui peuvent surprendre, comme par exemple *chef*, fortement positif. Ils correspondent à l'utilisation d'expressions (en l'occurrence *chef d'oeuvre*) et montrent l'insuffisance de l'approche par lemme.
– Enfin, les scores erronés observés dans les corpus de livres et de films sont fréquemment dus au fait qu'une partie du texte traite du contenu de l'œuvre. Les termes a priori dépréciatifs que l'on peut trouver dans ces descriptions ne préjugent en rien de l'opinion portée au film ou au livre. Ces observations incitent à développer un système qui permette de sélectionner les passages du texte qui correspondent effectivement à un avis donné sur le produit plutôt qu'à sa simple description.

Nous avons mis en œuvre une analyse de surface (*chunking*) qui détecte les groupes nominaux et verbaux et reconnaît certaines expressions figées. L'étude se réduit actuellement à la prise en compte des négations explicites à l'intérieur des groupes verbaux. Bien qu'embryonnaire, ses résultats (donnés dans la dernière colonne du tableau 1) nous incitent à poursuivre ce travail. Cependant, si une approche utilisant une analyse de surface semble prometteuse, son défaut est qu'elle est, bien plus encore qu'une approche en sac de mots, sensible à la qualité de la langue et en particulier, à l'orthographe. L'un de nos travaux en cours consiste à améliorer le prétraitement des corpus pour corriger, au moins partiellement, ce type de problème.

L'objectif est de permettre de dépasser une approche qui consiste à réduire les textes en un ensemble de mots en oubliant que c'est leur association qui fait sens. Reste à expérimenter si le bruit introduit permettra de dépasser l'efficacité du simple et robuste *sac de mots*.

Références

BRECK E. & CARDIE C. (2016). *Oxford Handbook of Computational Linguistics.*, chapter Opinion mining and sentiment analysis. Oxford University press, 2nd edition. *to appear.*

DADVAR M., HAUFF C. & DE JONG F. *(2011). Scope of negation detection in sentiment analysis. In* Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011*, p. 16–19, Amsterdam : University of Amsterdam. ISBN=not assigned.*

FELDMAN R. *(2013). Techniques and applications for sentiment analysis.* Commun. ACM*,* **56***(4), 82–89.*

FERRAND L. *(2001). Normes d'associations verbales pour 260 mots abstraits.* L'Année Psychologique*,* **101***, 683–721.*

FERRAND L. & ALARIO F. *(1998). Normes d'associations verbales pour 366 noms d'objets concrets.* L'Année Psychologique*,* **98***, 659–709.*

HOGENBOOM A., VAN ITERSON P., HEERSCHOP B., FRASINCAR F. & KAYMAK U. *(2011). Determining negation scope and strength in sentiment analysis. In* Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, October 9-12, 2011*, p. 2589–2594 : IEEE.*

JIA L., YU C. & MENG W. *(2009). The effect of negation on sentiment analysis and retrieval effectiveness. In* Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 1827–1830, New York, NY, USA : ACM.*

KAMPS J., MARX M., MOKKEN R. J. & DE RIJKE M. *(2004). Using WordNet to Measure Semantic Orientations of Adjectives. In* In Proceedings of LREC-04, 4th international conference on language resources and evaluation, Lisbon, PT*, volume 4, p. 1115–1118.*

PANG B. & LEE L. *(2008). Opinion mining and sentiment analysis.* Found. Trends Inf. Retr.*,* **2***(1-2), 1–135.*

QIU G., LIU B., BU J. & CHEN C. *(2011). Opinion word expansion and target extraction through double propagation.* Comput. Linguist.*,* **37***(1), 9–27.*

SYSSAU A. & FONT N. *(2005). évaluations des caractéristiques émotionnelles d'un corpus de 604 mots.* Bulletin de psychologie*,* **3***(477), 361–367.*

VINCENT M. & WINTERSTEIN G. *(2013). Construction et exploitation d'un corpus français pour l'analyse de sentiment. In* TALN-RÉCITAL 2013*, Les Sables d'Olonne, France.*

VINCZE N. & BESTGEN Y. *(2011). Une procédure automatique pour étendre des normes lexicales par l'analyse des cooccurrences dans des textes.* TAL*,* **52***(3), 191–216.*

WESTERSKI A. *(2007). Sentiment analysis : Introduction and the state of the art overview.* Universidad Politecnica de Madrid*, p. 211–218.*

WIEGAND M., BALAHUR A., ROTH B., KLAKOW D. & MONTOYO A. *(2010). A survey on the role of negation in sentiment analysis. In* Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, p. 60–68, Stroudsburg, PA, USA : Association for Computational Linguistics.*

WILSON T., WIEBE J. & HOFFMANN P. *(2005). Recognizing contextual polarity in phrase-level sentiment analysis. In* Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 347–354, Stroudsburg, PA, USA : Association for Computational Linguistics.*

ZHANG L., FERRARI S. & ENJALBERT P. *(2012). Opinion analysis : The effect of negation on polarity and intensity. In* J. JANCSARY*, Ed.,* Proceedings of KONVENS 2012*, p. 282–290 : ÖGAI. PATHOS 2012 workshop.*