Intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS

Wejdene Khiari *,**,***, Mathieu Roche ***, Asma Bouhafs Hafsia **

* Ecole Supérieur de Commerce de Tunis (ESC), Manouba, Tunisie
wijdenkhiari@gmail.com
** Institut de Hautes Etudes Commerciales(IHEC), Carthage, Tunisie
asma _ bouhafs@yahoo.com
*** TETIS, Cirad, Irstea, AgroParisTech & LIRMM, CNRS, Univ. Montpellier – France
mathieu.roche@cirad.fr

Contexte. Pendant la dernière décennie, Internet a connu une plus vaste portée grâce à l'émergence du Web social 2.0. Ceci a conduit au développement de nouveaux médias tels que les réseaux sociaux et à des travaux associés sur l'analyse des sentiments. L'objectif de cet article est de proposer une méthode de détection automatique de sentiments à partir du corpus 88milSMS (http://88milsms.huma-num.fr) en prenant en considération les spécificités de l'écriture SMS (Panckhurst et al., 2013). Dans ce cadre, nous nous intéressons à l'intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS.
Méthode. Dans un premier temps, nous avons identifié les SMS possédant des mots avec allongements à partir de trois caractères. Puis, de tels mots sont cherchés afin de construire un dictionnaire des mots et des mots allongés associés (exemple, merci / merciiii, merciiiii, merciiiiiii). Dans un deuxième temps, nous avons constitué un échantillon représentatif de 304 SMS possédant des mots allongés et 182 SMS sans allongement. Ce corpus a été annoté manuellement suivant l'opinion véhiculée : "Très positif", "Positif ", "Négatif", "Très négatif", "Neutre", "Je ne sais pas". Nous présentons ensuite une méthode fondée sur l'apprentissage supervisé qui s'appuie sur la représentation vectorielle des SMS sous forme de "sacs de mots" (Salton et al., 1975). Une représentation booléenne peut alors être effectuée sur la base des vecteurs relatifs à chaque SMS. Enfin, nous avons utilisé un lexique de sentiments et d'émotions FEEL (Abdaoui et al., 2014) pour pondérer certains mots porteurs de sentiments. Nous avons considéré que si un mot est présent dans ce dictionnaire, l'attribut correspondant est instancié à 2 dans la représentation vectorielle. Par exemple, si le mot "besoin" est présent dans le dictionnaire d'opinion, l'attribut est alors instancié à 2 dans les SMS. Si un attribut est présent dans un SMS, mais absent du dictionnaire, la valeur est instanciée à 1. En absence du mot dans le SMS, la valeur 0 est introduite. Et si un mot allongé est présent dans le dictionnaire d'opinion sous sa forme désallongée, l'attribut correspondant est instancié à 4 dans les SMS (vecteurs) comportant ce mot. Par exemple, si le mot allongé "besoinnnnn" est présent dans le dictionnaire d'opinion sous sa forme désallongée comme "besoin" l'attribut est instancié à 4.
Résultats. Nous avons procédé à une série d'expérimentations sur les différents jeux de données présents dans le Tableau 1 : (1) les corpus "SMS allongés" et (2) "SMS non allongés", (3) le corpus "SMS allongés Dico" issu de l'intégration du dictionnaire d'opinion et des SMS allongés, (4) le corpus "SMS non allongés Dico" issu de l'intégration du dictionnaire d'opinion

et des SMS non allongés, (5) le corpus "SMS désallongés" pour lequel nous avons supprimé la répétition de caractères des mots possédant un allongement. Les résultats obtenus en termes d'exactitude (accuracy) avec les algorithmes [1] selon 10-validation croisée à partir du logiciel Weka (Hall et al., 2009) sont donnés dans le Tableau 1. D'après cette analyse, nous remarquons que les SMS non allongés sont toujours mieux classés que les SMS allongés. Le fait d'appliquer un processus de "désallongement" des mots permet d'améliorer les résultats (30.26% vs 45.39% pour SMO et 29.60% vs. 41.77% pour J48).

	SMO	J48		SMO	J48
(1) SMS allongés	30.26	29.60	(3) SMS allongés Dico	50.65	46.38
(2) SMS non allongés	46.15	47.25	(4) SMS non allongés Dico	**64.48**	**64.48**
			(5) SMS désallongés	45.39	41.77

TAB. 1 – *Résultats fournis en termes d'exactitude (accuracy).*

Conclusion. Dans cet article, nous avons mis en place une nouvelle méthode pour la détection automatique de sentiments à partir d'un corpus de SMS réputé difficile à traiter. Notre contribution est d'identifier les descripteurs linguistiques qui véhiculent les opinions afin de proposer un modèle adapté à l'analyse des sentiments dans les SMS. Comme perspectives, nous prévoyons tester plusieurs autres algorithmes et envisager d'appliquer d'autres pondérations statistiques pour représenter les données textuelles par une représentation de type TF-IDF.

Références

Abdaoui, A., J. Azé, S. Bringay, et P. Poncelet (2014). Feel : French extended emotional lexicon. *ELRA Catalogue of Language Resources. ISLRN : 041-639-484-224-2*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : An update. *SIGKDD Explor. Newsl 11*(1), 10–18.

Panckhurst, R., C. Détrie, C. Lopez, C. Moïse, M. Roche, et B. Verine (2013). Sud4science de l'acquisition d'un grand corpus de sms en français à l'analyse de l'écriture sms. *Épistémé - revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques*.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(8), 613–620.

Summary

With the explosive growth of the social media (forums, blogs, and social networks) on the Web, the exploitation of these new information sources became essential. In this paper, we present a new automatic method to integrate knowledge for sentiment detection from a SMS corpus by combining lexical and semantic information.

1. Algorithmes appliqués avec les paramètres par défaut de Weka, par exemple le noyau polynomial pour SMO, la méthode à base d'arbre de décision J48.