# 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