Comment prendre en compte les spécificités de "l'écriture SMS" pour l'analyse de sentiments ?

Wejdene Khiari[1, 2, 3], Asma Bouhafs[2], Mathieu Roche[3]

[1]Ecole Supérieur de Commerce de Tunis (ESC), Manouba – Tunisie

wijdenkhiari@gmail.com

[2]Institut de Hautes Etudes Commerciales (IHEC), Carthage – Tunisie

asma _ bouhafs@yahoo.com

[3]TETIS, Cirad, Irstea, AgroParisTech & LIRMM, CNRS, Univ. Montpellier – France

mathieu.roche@cirad.fr

Abstract

With the explosive growth of textual data from social media (forums, blogs, and social networks), exploitation of these new information sources has become crucial. Our work focuses on sentiment analysis in this social media context. In order to identify sentiments in messages (e.g., tweets, SMS), original text-mining techniques have to be proposed. This paper presents a new method that integrates semantic and lexical knowledge for sentiment analysis. The proposed approach gives an important weight to "sentiment words" for a classification task. Our study compares two corpora (i.e. 88milSMS and DEFT'2015) in order to highlight the specific characteristics of SMS data in social media context.

Résumé

Avec la croissance explosive des données textuelles issues des médias sociaux (forums, blogs et réseaux sociaux), l'exploitation de ces nouvelles sources d'information est devenue cruciale. Nos travaux se concentrent sur l'analyse des sentiments dans ce contexte des médias sociaux. Pour identifier des sentiments issus des messages (par exemple, tweets, SMS), des techniques originales de fouille de textes doivent être proposées. Cet article présente une nouvelle méthode qui intègre les connaissances sémantiques et lexicales pour l'analyse des sentiments. L'approche proposée donne un poids important aux "mots de sentiment" pour une tâche de classification. Notre étude compare deux corpus (88milSMS et DEFT'2015) afin de mettre en évidence les caractéristiques spécifiques des données SMS dans le contexte des médias sociaux.

Mots clés : Traitement Automatique du Langage (TAL), fouille de textes, détection d'opinion, analyse de sentiments, corpus 88milSMS, informations lexicales et sémantiques.

1. Introduction

Pendant la dernière décennie, Internet a connu une plus vaste portée grâce à l'émergence du Web social (Web 2.0). Ceci a conduit au développement de nouveaux médias tels que les réseaux sociaux. Ces derniers sont multiples, nous citons par exemple, Twitter, Facebook, Google+ et LinkedIn. Ces sites Web offrent les possibilités aux utilisateurs d'exprimer et échanger des opinions et des idées avec les autres par le biais de multiples plateformes telles que les microblogues, blogues, sites Web, SMS, courriels, etc.

L'analyse automatique de textes issus de ces modes de communication pour la détection de sentiments est un vrai défi dans le domaine de la fouille d'opinions. Dans cet article, nous ne

distinguons pas réellement les concepts de « sentiment » (expression affective-intellective) et d'« opinion » (expression intellective).

Nos travaux s'inscrivent dans cette voie et concerne la détection de sentiments dans la communication écrite médiée. Nous nous intéressons à une tâche particulière de fouille de données qui porte sur la détection de sentiments à partir d'un corpus d'écrits non standards (messages SMS) sur lequel nous concentrons nos efforts. Le projet sms4science est un projet international coordonné par le CENTAL, un centre de recherche de l'Université Catholique de Louvain (Belgique).

Dans ce large contexte, l'objectif du projet sud4science (Panckhurst et al., 2013) est d'effectuer des recherches pluridisciplinaires sur un corpus de 88.522 SMS authentiques, en langue française, recueilli en 2011 puis anonymisé[1] (corpus 88milSMS disponible : http://88milsms.huma-num.fr).

Dans cet article, nous présentons une méthode de détection automatique de sentiments à partir du corpus 88milSMS en prenant en considération les spécificités de l'écriture SMS. Dans ce cadre, nous nous intéressons à l'intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS. Notre travail s'intéresse à un corpus très particulier et très difficile. Le corpus disponible le plus proche est celui issu de DEFT'2015. Nous nous sommes plus spécifiquement concentrés sur la tâche 1 qui s'intéresse à la classification de messages courts en français (tweets) selon leur polarité.

L'article est structuré de la façon suivante. La section 2 décrit les travaux relatifs à l'identification de sentiments dans des textes courts de type SMS et tweets. Dans la section 3, nous détaillons la méthodologie globale de notre processus d'intégration de connaissances. Dans la section 4, nous réalisons des expérimentations sur les différents jeux de données avec une étude comparative avec le corpus DEFT afin de valider la généricité de notre méthode. Enfin, dans la section 5, nous concluons en présentant de nouvelles perspectives à ces travaux.

2. État de l'art

Le domaine de l'analyse de sentiments également appelée fouille d'opinions a connu un intérêt croissant depuis le début des années 2000 (e.g. (Boiy et al., 2007), (Strapparava et Mihalcea, 2008), (Liu, 2012)) avec le développement du Web social et collaboratif 2.0, qui a favorisé l'émergence des réseaux sociaux.

Twitter est le site de microblogage le plus utilisé, avec environ 500 millions d'utilisateurs et 340 millions de tweets par jour. Il permet aux utilisateurs de publier des tweets de 140 caractères au maximum et de lire les messages des autres utilisateurs.

L'analyse des sentiments sur Twitter a attiré beaucoup d'attention récemment. Amir et al. (2014) décrivent leur participation à la tâche de classification de polarité de message à SemEval 2014 dans l'analyse des sentiments sur Twitter. La tâche de classification consiste à déterminer la polarité d'un message (positif, négatif ou neutre). Hangya et al. (2013) proposent également une méthode d'apprentissage supervisée fondée sur les unigrammes qui est appliquée sur les messages courts comme les tweets. L'objectif est de construire des

[1] Le but de l'anonymisation est de masquer l'identité d'un individu. Les étiquettes utilisées pour l'anonymisation sont les suivantes : Prénom (PRE), Nom (NOM), Surnom (SUR), Adresse (ADR), Lieu (LIE), Numéro de téléphone (TEL), Code (COD), URL (URL), Marque (MAR), Courriel (MEL), Autre.

modèles qui classent les tweets en trois catégories (positives, négatives ou neutres) par leur contenu. Pour déterminer la polarité d'un mot, les auteurs utilisent le lexique de sentiment SentiWordNet (Esuli et Sebastiani, 2006). Cette ressource est un lexique d'opinion dérivé de la base de données WordNet où chaque terme est associé à des scores numériques indiquant les informations (polarité, intensité) liées aux sentiments. Taboada et al. (2011) utilisent, quant à eux, un lexique pour extraire les mots véhiculant des sentiments (y compris les adjectifs, verbes, noms, et adverbes) dans un texte en combinant l'utilisation de corpus et de dictionnaires avec l'application d'une calculatrice d'orientation sémantique (SO-CAL). Cette dernière s'appuie sur des dictionnaires de mots annotés avec leur orientation sémantique et intègre l'intensité et la négation. L'évaluation de cette approche a montré que la méthode d'analyse de sentiment à base de lexique aboutit à une bonne performance et peut être facilement améliorée avec de multiples sources de connaissances. Fernández et al. (2014) proposent une approche supervisée dans l'analyse automatique de sentiment sur Twitter pour déterminer si un message exprime un sentiment positif, négatif ou neutre dans l'objectif de créer un classificateur de polarité fiable. La nouveauté de cette approche réside dans l'utilisation non seulement des mots et des n-grammes mais également des skipgrams comme descripteurs. La technique "skipgrams" est largement utilisée dans le domaine du traitement de la parole dans lequel des n-grammes étendus (n-grammes avec décalages) sont formés. La méthode des n-grammes est utilisée pour identifier les ensembles formés de n caractères consécutifs de différentes chaînes de caractères. Enfin, sur la base de motivations similaires à nos travaux, des approches s'intéressent à l'influence des répétitions de caractères pour la détection d'opinions (Brody et Diakopoulos, 2011), (Giachanou et Crestani, 2016). Ces approches s'appuient sur l'utilisation de lexiques (Brody et Diakopoulos, 2011) et l'intégration de figures stylistiques dans une "mesure d'opinion" (Giachanou et Crestani, 2016).

D'autres approches ont été proposées et se basent sur l'étude de corpus de SMS. « Short Message Service » ou « service de messages succincts » est un service qui permet aux usagers d'envoyer ou de recevoir des messages alphanumériques courts (moins de 160 caractères). Les travaux de Cougnon (2008) se fondent sur l'étude d'un corpus de 30.000 SMS accompagné d'un logiciel de consultation. L'objectif de cette recherche est l'étude des spécificités lexicales, phonétiques, morphologiques, voire syntaxiques (abréviations, spécialisation, répétition de caractères, allongement, etc.) pour la détection d'opinion ou l'analyse de sentiment dans les SMS. Cougnon et Thomas (2010) étudient la représentativité du corpus francophone de 30.000 SMS du Central en effectuant une série de tests du khi deux pour chaque dimension (âge, sexe, région d'origine, etc.).

Certains chercheurs travaillent sur la normalisation des SMS en graphie standard (Kobus et al., 2008), (Beaufort et al., 2008). D'autres se sont consacrés au respect de la norme écrite dans les SMS à travers trois phénomènes qui sont l'abréviation, les salutations et l'emprunt (Cougnon et Thomas, 2010). Les tests statistiques révèlent un important écart par rapport à la norme chez les utilisateurs. Le premier type d'écart est l'emploi des formes de salutations qui appartient au registre familier comme (kikou, hep, hello) qui ne représente que 26% des occurrences. Le second type est l'emploi des formes courantes de salutations (par ex. bonjour, salut, coucou et hello) qui sont très employées dans le corpus de SMS. Alors que, les messages qui ne sont pas abrégés représentent seulement 20% des données étudiées. Fernández et al. (2014) se sont basés sur la normalisation de chaque tweet. Ceci est réalisé par la conversion en minuscules de tous les caractères du texte (tweet), l'élimination des caractères répétés en considérant que si le même caractère est répété plus de 3 fois les répétitions suivantes sont supprimées, la substitution des noms d'utilisateurs et de mot-dièses

(hashtag en anglais) par les expressions USERNAME et HASHTAG. (Hangya et al., 2013) indiquent que la détection de la polarité d'un tweet n'est possible que si les étapes de normalisation sont appliquées : conversion de tous les mots en minuscules, remplacement des balises spécifiques de Twitter comme @ et # par les notations [USER] et [TAG], regroupement des émoticônes en classes positives et négatives. Par ailleurs, les caractères redondants sont supprimés dans le cas de mots contenant le même caractère au moins trois fois de suite, la conversion de toutes les URL dans les messages par l'annotation [URL] est effectuée, la conversion des chiffres sous la forme [NUMBER] et le remplacement des points d'interrogation et des points d'exclamation avec les notations [QUESTION MARK] et [EXCLAMATION MARK] sont également appliqués.

3. Processus d'intégration de connaissances

L'état de l'art met en relief qu'il existe peu de travaux en TAL qui étudient l'influence des allongements (nombre de répétitions, type de caractères répétés, etc.) dans les SMS. La nouveauté de notre approche est qu'elle permet d'étudier l'influence des aspects lexicaux et sémantiques propres aux SMS. L'objectif de cette section est de présenter notre méthode automatique de détection de sentiments que nous avons utilisée. Le processus général est décrit en Figure 1. Notre approche se décompose en quatre phases détaillées ci-après.

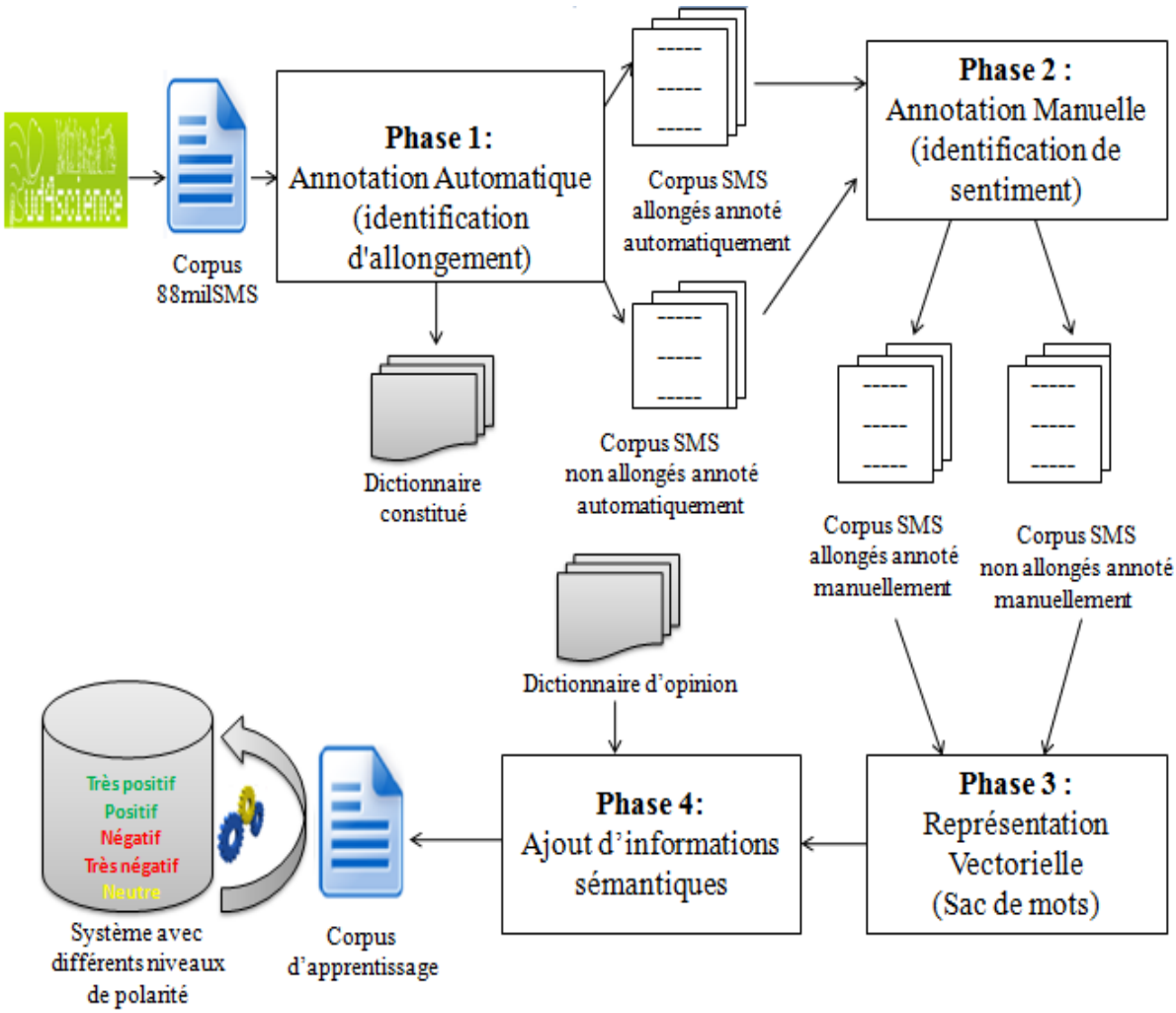

Figure 1 : Processus d'intégration de connaissances.

3.1. Phase 1: Annotation automatique

Pour l'annotation de corpus, nous avons commencé par isoler tous les SMS possédant des allongements, c'est-à-dire des répétitions de caractères à partir du corpus 88milSMS. Nous avons trouvé 14.209 SMS allongés. Nous avons choisi d'isoler tous les SMS avec la répétition des 5 voyelles (a, e, i, o, u), 5 consonnes (g, r, t, c, d) en majuscules et minuscules à partir de trois caractères consécutifs comme les travaux de Fernández et al. (2014) et les points d'exclamation. Par exemple, si un mot contient le même caractère plus de trois fois de manière consécutive, la séquence est isolée. L'exemple suivant (cf. Tableau 1) montre des SMS allongés qui sont isolés automatiquement.

Num SMS	*SMS*
26526	Merciiiiiii ... Je prendrai 2 heures de pause ... [Merciiiiiii]
46157	J'ai tt essayer facebook veut paaaaaaas [paaaaaaas]
50764	Alors alors alors ? Biiiiisous [Biiiiisous]

*Tableau 1 **:** Exemple d'annotation automatique des SMS allongés. Les éléments entre crochets sont relatifs aux mots allongés identifiés.*

À partir des mots allongés, nous avons recherché les SMS possédant ces mêmes mots non allongés (entre crochets nous avons ajouté les mots allongés associés) qui constituent la deuxième partie du corpus (cf. Tableau 2).

Num SMS	*SMS*
12140	Merci ptit < PRE _ 3 > [Merciiii, Merciiiiiii]
12199	Jv essayer je promet pas. [paaaaaaas, paaaaas, paaas]
12132	Merci bisou [Merciiii, Merciiiiiii, bisouuus, biiiisous, biiiiisous, biiisous]

Tableau 2 : Exemple d'annotation automatique des SMS non allongés.

En conséquence, nous avons constitué un dictionnaire des mots de manière similaire aux travaux de (Brody et Diakopoulos, 2011). Ce dictionnaire contient tous les mots et les allongements possibles retrouvés dans le corpus (cf. Tableau 3).

Mots	*Liste des allongements associés*
merci	merciiii, merciiiii, merciiiiiii
demain	demaiiin, demaaaaaaain, demaaain
pas	paaaaaaas, paaaaas, paaas, paaaas, paaaaaas, paaaaaaaas
quoi	quoiiii, quoiiiiiiii, quoooooii, quooooooooooiiiiiiiiii

Tableau 3: Quelques exemples de mots avec une liste des allongements associés extraits à partir du corpus 88milSMS.

3.2. Phase 2 : Annotation manuelle

Dans un deuxième temps, nous avons réalisé une synthèse afin de répertorier le nombre d'allongements sur le corpus des SMS allongés et des SMS non allongés (cf. Phase 2 de la Figure 1). En particulier, nous avons recherché les allongements d'une taille précise de 3, 4 et plus de 4 pour les 5 voyelles (a, e, i, o, u), les 5 consonnes (g, r, t, c, d) et les points d'exclamation. Ces données ont été triées selon un ordre alphabétique pour constituer un corpus de 5222 SMS possédant des allongements. Ensuite, nous avons constitué un échantillon représentatif de 304 SMS possédant des mots allongés et 182 SMS sans allongement. Par la suite, ce corpus a été annoté manuellement. Notre but est de pouvoir identifier les SMS selon le sentiment qu'ils expriment.

Pour ce faire, nous avons d'abord constitué un corpus d'apprentissage. L'identification des sentiments nécessaire pour former le corpus d'apprentissage est effectuée manuellement. Nous déterminons le sentiment contenu dans le message par une polarité allant de (i) 5 pour un SMS qui exprime une opinion très positive à (ii) 4 s'il s'agit d'un SMS portant une opinion positive et allant de (iii) 2 pour un SMS qui exprime une opinion très négative à (iv) 3 s'il s'agit d'un SMS qui peut être associé à une opinion négative. Un SMS neutre est annoté à 1 alors qu'un SMS que nous ne pouvons "polariser" est annoté à 0. Les Tableaux 4 et 5 présentent des exemples de SMS allongés et de SMS non allongés qui sont annotés manuellement suivant les 6 catégories d'opinions avec la polarité résultante. Notons que la polarité de certains SMS peut se révéler complexe à identifier (par exemple, "Momooooooon !") alors que d'autres sont plus aisés à analyser (par exemple, "Je taaaaaime").

SMS	Polarité
Je taaaaaime	5
Mdrrr ah ces bon souvenir xD	4
Je m'ennuiiie	3
Putain, ton scenar est voué à l'echec pour une seule et unique raison tellement nuuuuulle. T'as pas numeroté les pages PETIT BOL DE MERDE !	2
Momooooooon !	1
Gnagnaaandmtgmpdtwamdgdavngd <3333	0

Tableau 4 : Exemples d'annotation manuelle des SMS allongés.

SMS	Polarité
Non, je suis à la soirée de mes parents. Je te fais de gros bisous, je t'aime très fort. Je t'appelle demain	5
:) aller courage	4
Nn <PRE 3 > elle est trop chiante.	3
Ahh putain la chance ! X) mais bon si tu viens a 9h c'est dla merde	2
Oui je vois	1
10 ^^ '	0

Tableau 5 : Exemples d'annotation manuelle des SMS non allongés.

3.3. Phase 3 : Représentation vectorielle du corpus

Une fois le corpus annoté constitué, ce dernier sera traduit en suivant une représentation vectorielle des textes (Salton et al., 1975). Dans cette phase (cf. Phase 3 de la Figure 1), nous présentons une méthode fondée sur l'apprentissage supervisé.

Un traitement préalable peut consister à éliminer les messages classés comme "je ne sais pas" pour avoir les 5 catégories de classe d'opinions. Les données sont représentées sous forme vectorielle, où chaque SMS est représenté par un vecteur d'attributs. Ces derniers peuvent être les mots, les n-grammes de mots, les n-grammes de caractères, la ponctuation, les émoticônes, la négation et les mots allongés. Le dernier attribut correspond à la classe de sentiment.

Une représentation booléenne peut alors être effectuée sur la base des vecteurs relatifs à chaque SMS (présence ou absence des descripteurs dans les SMS).

3.4. Phase 4 : Ajout d'informations sémantiques

Dans la dernière phase (cf. Phase 4 de la Figure 1), nous avons ajouté des informations sémantiques issues de dictionnaires dédiés (dictionnaires d'opinion) avec le lexique français FEEL[2] des émotions et sentiments de (Abdaoui et al., 2014).

Ce lexique est formé de manière semi-automatique en traduisant le lexique de sentiments et d'émotions anglais NRC-Canada (Mohammad et Turney, 2010). Il est composé de plus de 14.000 mots distincts selon leur polarité et associés à 6 émotions de (Ekman, 1992) : confiance, peur, tristesse, colère, surprise, dégoût, joie. La traduction semi-automatique supervisée de ce lexique est réalisée par un traducteur humain expérimenté. Il a été étendu en anglais et en français par l'étude des synonymes et des antonymes qui sont validés en termes d'impact pour une tâche de classification automatique pour les deux types de sentiment (polarité et émotion) et différents jeux de données classiques de la littérature. Notons que chaque mot peut être associé à plusieurs émotions.

Par exemple, le mot "merci" présent dans ce lexique est associé à l'émotion "confiance" et "joie" et sa polarité est positive.

Pour cela, nous avons commencé par effectuer la fusion des catégories (i.e positive et très positive / négative et très négative) pour obtenir trois classes : classe positive, négative et neutre. Ce processus de fusion est motivé par le fait que la distinction entre les classes très positives et positives (resp. négatives et très négatives) est très subtile et discutable.

Ensuite, nous avons intégré les informations propres à ce dictionnaire d'opinion aux SMS pour construire deux corpus d'apprentissage : corpus "SMS allongés Dico" issu de l'intégration du dictionnaire d'opinion aux SMS allongés et corpus "SMS non allongés Dico" issu de l'intégration du dictionnaire d'opinion aux SMS non allongés afin de pondérer certains mots porteurs de sentiments.

Pour la représentation vectorielle, nous avons réalisé plusieurs types de pondérations, nous avons pondéré différemment selon le nombre de caractères propres à l'allongement. Pour ce faire, nous proposons la pondération suivante pour chaque mot des SMS:

$$\text{SentiPond} = k * \alpha * \log(1 + c)$$

[2] https ://www.lirmm.fr/patient-mind/pmwiki/pmwiki.php?n=Site.Ressources

Avec :

- k = 0,1: si un mot est présent dans le message, k prend la valeur 1 sinon k prend la valeur 0.
- c est le nombre de caractères de l'allongement (allongement maximum qu'il y a dans le mot dans le cas de plus de 2 allongements),
- α = 2, 1 si le mot appartient ou non au dictionnaire d'opinion.

Exemple : si le mot allongé "besoinnnnn" est présent dans le SMS et dans le lexique d'opinions (dictionnaire), on aura ainsi :

- SentiPond = 1*2* log(1+5) = 1,5
- Si le mot n'appartient pas au dictionnaire, le poids du premier exemple (besoinnnnn) sera de 1*1* log(1+5) = 0.7

Notre motivation de ce choix est de donner plus de poids aux mots véhiculant un sentiment (dictionnaire et allongement) et de prendre en compte les aspects sémantiques et lexicaux propres aux SMS. Par exemple les répétitions de caractères, de phonèmes ou de marques de ponctuation (adorableeeeee, riiiiiche) sont souvent porteuses de sentiment que notre pondération permet de privilégier. Notre but est la constitution d'un corpus d'apprentissage pour pouvoir apprendre un modèle. Ceci permettra de prédire la polarité des SMS avec différents niveaux de polarité.

Outre les expérimentations sur le corpus 88milSMS, par la suite, nous allons tester notre approche à partir d'un corpus de tweets issus de DEFT'2015. Ce défi est un atelier annuel d'évaluation francophone en fouille de textes qui a porté sur l'analyse de l'opinion, des sentiments et des émotions dans des tweets rédigés en français. Le défi de l'édition 2015 propose trois tâches, la première tâche (tâche 1) vise à classer les messages selon leur polarité (positive, négative ou neutre) (cf. Tableau 6) pour former le corpus "T1 DEFT 2015" qui possède les caractéristiques suivantes : nombre d'attributs : 2317 + classe, nombre d'instances : 7869, type : numérique, nombre de classes : 3.

Tweets	*Polarité*
Je tiens à féliciter les écologistes qui ont réussi à surpasser, dans mon échelle d'écœurement, les nutritionnistes.	=
Rain forest alliance est plus axé sur le développement durable, l'écologie #arte	+
Pelouses sèches : un réservoir de biodiversité aujourd'hui menacé: Bien que reconnues comme habitat d'intérêt ... http://t.co/G9NEd5Pbis	-

Tableau 6 : Exemples de tweets avec des polarités.

4. Expérimentations

Dans cette section, nous présentons les résultats de l'évaluation de notre méthode. Les expérimentations ont été réalisées sur deux bases (corpus) : une base de SMS allongés et une base de SMS non allongés. Les données sont stockées dans ces deux bases au format ARFF

(Attribute Relationship File Format) tel que requis par l'environnement Weka (Hall et al., 2009). La pondération SentiPond a été appliquée. Les Tableaux 7 et 8 présentent les caractéristiques des corpus expérimentés.

	Nombre d'instances	*Nombre d'attributs*	*Nombre de classes*
SMS allongés	304	2053	3
SMS non allongés	182	1470	3

Tableau 7: Caractéristiques des corpus expérimentés.

Classes d'opinion	5	4	3	2	1
SMS allongés	62	62	62	62	56
SMS non allongés	39	39	39	26	39

Tableau 8 : Nombre de SMS par classe avant la fusion des classes "positive" et "très positive" (resp. les classes "négative" et "très négative").

Sur chacun de ces corpus, nous avons exécuté 4 algorithmes[3] (SMO, J48, DMNB Text, Naive Bayes) à partir du logiciel Weka (Hall et al., 2009), les résultats en termes d'exactitude (accuracy) sont présentés dans le Tableau 9 (utilisation de la validation croisée à 10 échantillons).

	SMO	*J48*	*DMNB Text*	*Naive Bayes*
SMS allongés	50.54%	51.27%	51.27%	47.27%
SMS non allongés	63.62%	64.85%	63.62%	63.62%

Tableau 9 : Les résultats en termes d'exactitude (accuracy) en apprenant les modèles sur les corpus SMS allongés et SMS non allongés.

Nous remarquons (cf. Tableau 9) que les SMS non allongés sont beaucoup mieux classés que les SMS allongés avec un pourcentage d'instances correctement classées de 64% pour les SMS non allongés par rapport à 51% pour les SMS allongés dans les meilleures conditions.

Les algorithmes J48 et DMNBText possèdent le pourcentage d'exactitude le plus élevé lors de son application sur le modèle d'apprentissage "SMS allongés" uniquement. De plus, l'algorithme J48 donne un pourcentage d'exactitude plus élevé sur le modèle d'apprentissage

[3] 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, la classification bayésienne est utilisée comme une méthode d'apprentissage probabiliste pour DMNB Text et Naive Bayes.

SMS non allongés. Cependant, les résultats sont du même ordre avec les différentes méthodes de classification.

Notre deuxième expérimentation est réalisée en comparant les différents jeux de données présents dans le Tableau 10 : les bases "SMS allongés" et "SMS non allongés" (cf. Tableaux 7 et 8), le corpus "SMS allongés Dico" et "SMS non allongés Dico" (cf. Section 3.4), le corpus "SMS désallongés" pour lequel nous avons supprimé la répétition des caractères des mots possédant un allongement. Dans ce contexte, citons par exemple le mot allongé "Merciii" présent dans le fichier des SMS allongés devient dans le fichier des SMS désallongés sous la forme "Merci". Le corpus "T1 DEFT 2015" (cf. Section 3.4) est également considéré pour tester nos propositions.

	SMO	J48
SMS allongés	50.54	51.27
SMS non allongés	63.62	64.85
SMS allongés Dico	56.18	56.59
SMS non allongés Dico	**69.98**	**69.98**
SMS désallongés	51.67	50.77
T1 DEFT 2015	50.27	56.50

Tableau 10: Résultats fournis en termes d'exactitude (accuracy).

Comme le montre le Tableau 10, le système fournit de meilleurs résultats avec les classificateurs J48 et SMO (exactitude de 69.98%) appliqués au corpus "SMS non allongés Dico" de SMS non allongés. Enfin, nous obtenons une valeur d'exactitude égale à 56.50% pour le corpus "T1 DEFT 2015" (avec J48).

D'après cette analyse, nous remarquons que les SMS non allongés sont mieux classés que les SMS allongés. Le fait d'appliquer un processus de "désallongement" des mots n'améliore pas les résultats.

Dans nos travaux avec le corpus DEFT'2015, nous nous sommes concentrés sur un point particulier (intégration d'informations sémantiques et lexicales spécifiques) et non sur tous les aspects de l'analyse de sentiment comme les autres participants du défi. Les résultats obtenus sont cependant satisfaisants.

4. 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 (88milSMS). Nous avons mis en place un processus spécifique afin d'appliquer une méthode d'apprentissage supervisé. 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.

L'objectif de notre approche est de favoriser les descripteurs véhiculant un sentiment dans la représentation classique "Sac de mots". Pour ce faire, une pondération spécifique a été proposée pour les descripteurs selon leur caractéristique lexicale (présence d'un phénomène d'allongement) et/ou leur spécificité sémantique (présence de l'élément dans un dictionnaire dédié).

L'étude comparative avec le corpus de DEFT qui concerne la polarité de messages courts (tweets) en français a mis en évidence le caractère spécifique du corpus 88milSMS qui est tout à fait différent avec une création lexicale bien plus riche. Ainsi le corpus de SMS est beaucoup plus difficile à traiter que le corpus qui est constitué de tweets.

Bien que le travail effectué ait donné lieu à des résultats intéressants, nous envisageons dans les travaux à venir, de faire quelques modifications dans le but d'améliorer la performance de notre modèle. Comme perspectives, nous prévoyons de tester plusieurs autres algorithmes. Nous pourrons aussi envisager d'appliquer d'autres pondérations statistiques pour représenter les données textuelles par une représentation de type TF-IDF (et/ou OKAPI par exemple). Nous envisageons enfin d'approfondir l'étude des poids à adapter selon les caractéristiques lexicales et sémantiques des mots (allongés et non allongés).

Remerciements

Ces travaux sont financés par le projet SONGES - http://textmining.biz/Projects/Songes (FEDER et Languedoc-Roussillon). Notre étude fait suite au projet sud4science LR (http://sud4science.org/) coordonné par Rachel Panckhurst et soutenu par le CENTAL (Université Catholique de Louvain, Belgique), la MSH-M (Maison des Sciences de l'Homme de Montpellier) et la DGLFLF (Délégation générale à la langue française et aux langues de France).

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.

Amir, S., M. B. Almeida, B. Martins, J. a. Filgueiras, et M. J. Silva (2014). Tugas : Exploiting unlabelled data for twitter sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 673–677. Association for Computational Linguistics and Dublin City University.

Beaufort, R., S. Roekhaut, et C. Fairon (2008). Définition d'un système d'alignement sms français standard à l'aide d'un filtre de composition. In *Proceedings of JADT 2008*, pp. 155–166.

Boiy, E., P. Hens, K. Deschacht, et M. francine Moens (2007). Automatic sentiment analysis in on-line text. In *Proceedings of the 11th International Conference on Electronic Publishing*, pp. 349–360.

Brody S., N. Diakopoulos (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*. p.562-570.

Cougnon, L. A. (2008). Le français de belgique dans l'écrit spontané, approche syntaxique et phonétique d'un corpus de sms. Travaux du Cercle belge de linguistique.

Cougnon, L. A. et F. Thomas (2010). Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de sms. In *10th International Conference on statistical analysis of textual data (JADT 2010)*, pp. 9–11.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 169–200.

Esuli, A. et F. Sebastiani (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pp. 417–422.

Fernández, J., Y. Gutiérrez, J. M. Gómez, et P. Martinez-Barco (2014). Gplsi : Supervised sentiment analysis in twitter using skipgrams. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 294–299. Association for Computational Linguistics and Dublin City University.

Giachanou, A., F. Crestani (2016). Opinion Retrieval in Twitter Using Stylistic Variations. In *Proceedings of ACM Symposium on Applied Computing (SAC)*, p.1077-1079.

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.

Hangya, V., G. Berend, I. Varga, et R. Farkas (2013). Szte-nlp : Aspect level opinion mining exploiting syntactic cues. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, pp. 549–553. Association for Computational Linguistics.

Kobus, C., F. Yvon, et G. Damnati (2008). Normalizing sms : Are two metaphors better than one?. In *Proceedings of the 22Nd International Conference on Computational Linguistics Volume 1*, COLING 08, Stroudsburg, PA, USA, pp. 441–448. Association for Computational´ Linguistics.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Mohammad, S. M. et P. D. Turney (2010). Emotions evoked by common words and phrases : Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET 10, Stroudsburg, PA, USA, pp. 26–34. Association for Computational Linguistics.´

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.

Strapparava, C. et R. Mihalcea (2008). Learning to identify emotions in text. pp. 1556–1560. In SAC 08 Proceedings of the 2008 ACM symposium on Applied computing.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, et M. Stede (2011). Lexicon based methods for sentiment analysis. *Comput. Linguist. 37*(2), 267–307.