Integration of Lexical and Semantic Knowledge for Sentiment Analysis in SMS

Wejdene Khiari[*,**,***], **Asma Bouhafs**[**], **Mathieu Roche**[***]
[*]ESC Tunis School of Business, University of Manouba, Tunisia
[**]University of Carthage, Carthage Presidency, Tunisia
[***]UMR TETIS (Cirad, Irstea, AgroParisTech) & LIRMM (CNRS, Univ. Montpellier), France
wijdenkhiari@gmail.com, asma_bouhafs@yahoo.com, mathieu.roche@cirad.fr

Abstract

With the explosive growth of online social media (forums, blogs, and social networks), exploitation of these new information sources has become essential. Our work is based on the sud4science project. The goal of this project is to perform multidisciplinary work on a corpus of authentic SMS, in French, collected in 2011 and anonymised (88milSMS corpus: http://88milsms.huma-num.fr). This paper highlights a new method to integrate opinion detection knowledge from an SMS corpus by combining lexical and semantic information. More precisely, our approach gives more weight to words with a sentiment (i.e. presence of words in a dedicated dictionary) for a classification task based on three classes: positive, negative, and neutral. The experiments were conducted on two corpora: an elongated SMS corpus (i.e. repetitions of characters in messages) and a non-elongated SMS corpus. We noted that non-elongated SMS were much better classified than elongated SMS. Overall, this study highlighted that the integration of semantic knowledge always improves classification.

Keywords: Sentiment analysis, SMS corpus, lexical and semantic information.

1. Introduction

Internet has evolved boundlessly over the last decade with the advent of the social Web (Web 2.0). This has led to the development of new media such as various social networks ranging from Twitter, Facebook, Google+ and LinkedIn. These web sites offer opportunities for users to express themselves, as well as to exchange opinions and ideas with others through multiple platforms such as microblogs, blogs, web sites, SMS, emails, etc. Automatic analysis of texts generated from these communication modes for opinion detection is a real challenge in the field of opinion mining.

Our work is under way in this context. The sms4science project is coordinated by CENTAL (Centre for Natural Language Processing) at the Catholic University of Louvain, Belgium. The goal of the sud4science project (Panckhurst et al., 2013) is to perform multidisciplinary work on a corpus of 88.522 authentic SMS, in French, collected in 2011 and anonymised[1] (88milSMS corpus: http://88milsms.huma-num.fr).

In this paper, we present an opinion mining process that considers the specificities of SMS. The proposed paper is organized as follows. In Section 2, we describe our work related to sentiment identification in short texts, like those found in SMS and tweets. In Section 3, we detail the overall methodology of our knowledge integration process based on two strategies, automatic and manual, for annotation of the corpus. This annotation process confirmed the difficulties involved in sentiment analysis and feature identification in short texts. Then, we present a method based on supervised learning that relies on bag-of-words representation (Salton et al., 1975). In Section 4, we conduct experiments using the datasets to validate the performance of our method. Finally, Section 5 concludes and provides possible directions for future work.

2. Related work

Interest in the field of opinion mining and sentiment analysis has been growing since early 2000 (*e.g.* (Boiy et al., 2007), (Strapparava and Mihalcea, 2008), (Liu, 2012)) with the development of social and collaborative Web 2.0, which has favored the emergence of social networks.
Twitter is the most commonly used microblogging platform, with approximately 500 million users and 340 million tweets a day. It allows users to publish tweets of 140 characters at most and to read the messages of other users. Sentiment analysis on Twitter has drawn a lot of attention recently. (Amir et al., 2014) describe their participation in the message polarity classification task of SemEval 2014. The classification task consists of determining the polarity of a message (positive, negative, or neutral).
(Hangya et al., 2014) also propose a supervised learning method based on unigrams, which is applied to short messages like tweets. The goal is to build models that classify tweets into three classes according to their content. To determine the polarity of a word, they use the sentiment lexicon SentiWordNet (Esuli and Sebastiani, 2006). This resource is an opinion lexicon derived from the WordNet database in which each term is associated with numerical scores indicating the information (polarity, intensity) linked to the sentiment. (Taboada et al., 2011) use a lexicon to extract sentiment bearing words (including adjectives, verbs, nouns, and adverbs) of a text by combining the use of corpora and dictionaries with the application of a semantic orientation calculator (SO-CAL).

[1] The purpose of this anonymisation is to mask the identity of individuals (Panckhurst et al., 2013). The following tags were used for the anonymisation process: PRE (First Name), NOM (Last Name), SUR (Nickname), ADR (Address), LIE (Place), TEL (Telephone Number), COD (Code), URL (URL), MAR (Brand Name), MEL (Email), Other.

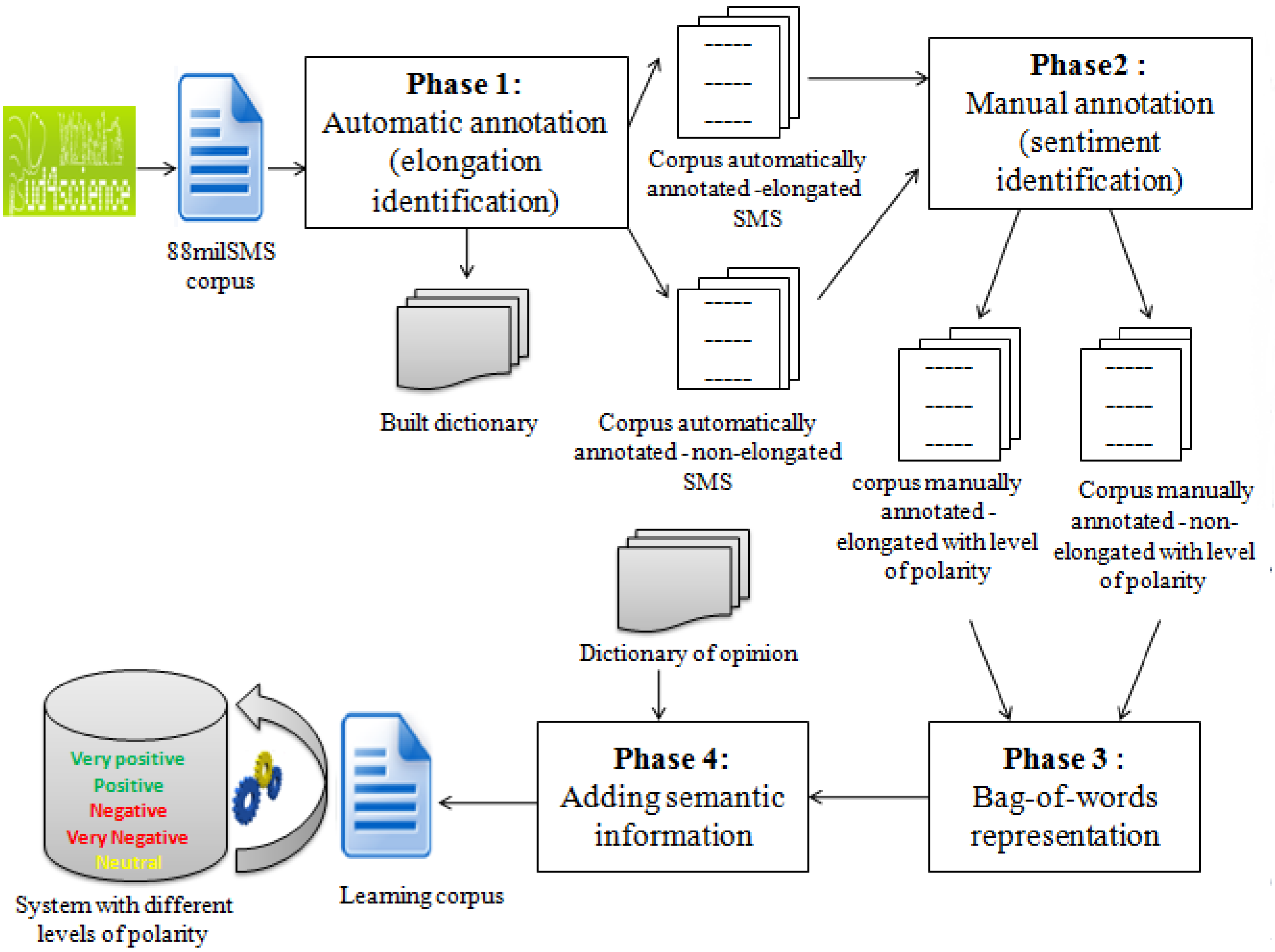
88milSMS corpus
Phase 1: Automatic annotation (elongation identification)
Built dictionary
Corpus automatically annotated-elongated SMS
Corpus automatically annotated-non-elongated SMS
Phase2 : Manual annotation (sentiment identification)
corpus manually annotated-elongated with level of polarity
Corpus manually annotated-non-elongated with level of polarity
Dictionary of opinion
Phase 3 : Bag-of-words representation
Phase 4: Adding semantic information
Learning corpus
Very positive
Positive
Negative
Very Negative
Neutral
System with different levels of polarity

Figure 1: Knowledge Integration Process.

Other approaches have been proposed and are based on the study of SMS corpora. The Short Message Service (SMS) allows users to send or receive short alphanumeric messages (less than 160 characters).
(Cougnon, 2008) conducted a study of a corpus of 30 000 SMS texts associated with consultation software. (Cougnon and Thomas, 2010) studied the representativeness of this corpus by performing statistical tests for different dimensions (age, sex, region of origin, etc.). They found that according to the sex and the age of users, 57.2% of women and 42.7% of men did not correspond to the 51.6% and 48.4% gender ratio within the overall population.
Some researchers have worked on the standardization of SMS to standard orthography (Kobus et al., 2008), (Beaufort et al., 2008). (Fernández et al., 2014) relied on the basic normalization of each tweet. This is done by converting all characters in the tweet text to lower case, eliminating the repetition of characters by considering that if the same character is repeated more than 3 times, the rest of the repetitions are removed.
(Hangya et al., 2014) indicated that detection of the polarity of a tweet is only possible if normalization steps are applied.

3. Knowledge Integration Processes

By our approach, we assessed the influence of lexical and semantic aspects for sentiment analysis in SMS. The general process is described in Figure 1. Our approach was divided into the four phases described below.

3.1. Phase 1: Automatic annotation

For corpus annotation, we began by isolating all SMS with elongations, i.e. repetitions of characters, from a sample of the 88milSMS corpus. We found 14 209 elongated SMS, and that the number of elongated SMS which we obtained is very large. We thus decided to isolate all SMS with the repetition of 5 vowels (a, e, i, o, u), 5 consonants (g, r, t, c, d) in upper and lower case from three consecutive characters as done in the study of (Fernández et al., 2014) and exclamation marks.
For example, if a word contained the same character more than three times consecutively, the SMS was isolated. The following example (see Table 1) shows a sample of elongated SMS, which were isolated automatically.
From the elongated words, we searched for SMS having the same words in a non-elongated form (we added the associated elongated words in square brackets) which constituted the second part of the corpus (see Table 2).

Num SMS	Content of SMS
5657	T’es paaaaas sur skype :([paaaaas]
7055	D’accooоord. [D’accoooord]
26526	Merciiiiiii ... Je prendrai 2 heures de pause ... [Merciiiiiii]
50764	Alors alors alors ? Biiiiisous [Biiiiisous]

Table 1: Example of automatic annotation of elongated SMS. Elements in square brackets are relative to the identified elongated words.

Num SMS	Content of SMS
19379	Je conai pas dsl [paaaaaaas,paaaaas,paaas,paaaas]
17163	Ah D’accord... [D’accoooord]
12140	Merci ptit < PRE _ 3> [Merciiii,Merciiiiiii]
23166	Des bisous [bisouuus,biiiisous,biiiiisous,biiisous]

Table 2: Example of automatic annotation of non-elongated SMS.

Accordingly, we constructed a dictionary of words associated with all possible elongations found in the corpus (see Table 3).

Words	List of associated elongations
faim	faiiiiiiim, faiiiim, faiiiiim, faiiim, faiiiiiim
faire	faiiiiiire
merci	merciiii, merciiiii, merciiiiiii
nuit	nuiiit, nuuuiiiiiit, nuuuuit
quoi	quoiiii, quoiiiiiiii, quoooоiii, quoooooooooiii-iiiiiii

Table 3: Some examples of words with a list of associated elongations extracted from the 88milSMS corpus.

3.2. Phase 2: Manual Annotation

Secondly, we computed statistics on the number of elongations in the corpus (see Phase 2 of Figure 1). In particular, we searched for the elongation of a specific size of 3, 4, and more than 4 for 5 vowels (a, e, i, o, u), 5 consonants (g, r, t, c, d), and exclamation marks. These data constituted a corpus of 5 222 SMS with elongations. Then we extracted a representative sample of 304 elongated SMS and 182 non-elongated SMS. We chose to work with this representative sample (see Tables 6 and 7 in Section 4) because when we manually appraised a representative sample of 522 SMS, we found an imbalance between classes.

Subsequently, this corpus was manually annotated. Our aim was to identify SMS retrieved according to the sentiment they expressed.

We thus constructed a learning corpus. Our aim was to determine the opinion contained in the messages according to a polarity ranging from: (i) 5 for an SMS expressing a very positive opinion, (ii) 4 in the case of an SMS with a positive opinion, (iii) 2 for an SMS expressing a very negative opinion, (iv) 3 for an SMS that could be associated with a negative opinion. A neutral SMS was annotated 1 while an SMS that we could not ”polarize” was annotated 0.

Tables 4 and 5 show examples of elongated and non-elongated SMS which were annotated manually according to the 6 categories of opinions with the resulting polarity.

SMS	Polarity
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
Momoooooоon!	1
Gnagnaaandmtgmpdtwamdgdavngd <3333	0

Table 4: Examples of manual annotation of elongated SMS.

SMS	Polarity
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

Table 5: Examples of manual annotation of non-elongated SMS.

3.3. Phase 3: Vector representation of the corpus

Once the annotated corpus was constituted, it was translated according to a vector representation of texts. This representation known as ”Salton” (Salton et al., 1975) or ”bag-of-words” is relatively effective for classification tasks. A Boolean representation was applied (presence or absence of features in SMS).

In this phase (see Phase 3 of Figure 1), we used a method based on supervised learning. A preliminary process consisted of eliminating messages classified as ”I do not know” so as to have 5 classes of opinions.

3.4. Phase 4: Adding semantic information

In the last phase (see Phase 4 of Figure 1), we added semantic information from an emotion dictionary [2] (Abdaoui et al., 2014). This lexicon contains more than 14 000 distinct words expressing emotions and sentiments according to their polarity and associated with 6 emotions of (Ekman, 1992). It was created by translating and expanding the

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

English Emotional Lexicon NRC-Canada (Mohammad and Turney, 2010). The process was supervised and validated manually by a human professional translator. It was extended in English and French by the study of synonyms and antonyms that are validated in terms of impact on an automatic classification task for the two types sentiment (polarity and emotion) and different classical datasets from the literature.

For this, we started by merging categories (i.e positive and very positive / negative and very negative) to obtain three classes: positive, negative, and neutral. We argued the relevance of this merging process because the distinction between very positive and positive classes (resp. negative and very negative) is very subtile and debatable.

We integrated the information related to this dictionary in order to build two learning corpora: (1) ”elongated SMS Dico” corpus obtained by integration of the opinion dictionary and elongated SMS, (2) ”non-elongated SMS Dico” corpus by integration of the opinion dictionary and non-elongated SMS.

We considered that if a word was present in the opinion dictionary, the corresponding attribute was instantiated at 2 in the vectorial representation. If the attribute was present in the SMS, but absent in the dictionary, the value was instantiated at 1. In the absence of the word in the SMS, the value 0 was introduced. And if an elongated word was present in the opinion dictionary in shortened form, the corresponding attribute was instantiated at 4 in SMS vectors. For example, if the elongated word ”besoinnnnn” was present in the opinion dictionary in its shortened form as ”besoin” the attribute was instantiated at 4.

By this choice, we wanted to give more weight to words with a sentiment (dictionary and elongation) while taking the semantic and lexical aspects into account. For example, repetitions of characters, phonemes or punctuation marks (e.g. Adorableeeeee, riiiiiche) were often bearers of sentiment that our weighting favored.

We aimed to compile a learning corpus in order to build a model to enable prediction of the polarity of SMS with various polarity levels.

4. Experiments

In this section, we present the results of the evaluation of our method. The experiments were conducted on two corpora: an elongated SMS corpus and a non-elongated SMS corpus. The data were stored in ARFF format (Attribute Relationship File Format) which is required for the Weka environment (Hall et al., 2009). Tables 6 and 7 present some characteristics related to our data.

On each of these corpora, we applied 4 algorithms (SMO, J48, DMNB Text, Naive Bayes) [3]. The results in terms of accuracy are presented in Table 8 using 10 cross-validations.

Corpus	Number of instances	Number of attributes	Number of classes
elongated SMS	304	2053	3
non-elongated SMS	182	1470	3

Table 6: Characteristics of our corpora.

Class opinion	5	4	3	2	1
elongated SMS	62	62	62	62	56
non-elongated SMS	39	39	39	26	39

Table 7: Number of SMS by class before merging ”positive” and ”very positive” classes (resp. ”negative” and ”very negative” classes).

Corpus	SMO	J48	DMNB Text	Naive Bayes
elongated SMS	**46.38**%	41.77%	46.05%	40.13%
non-elongated SMS	59.56%	**63.38**%	59.56%	52.45%

Table 8: Accuracy according to different algorithms and corpora.

We noted (see Table 8) that non-elongated SMS were much better classified than elongated SMS with 63% of instances correctly classified for non-elongated SMS compared to 46% for elongated SMS.

The SMO and DMNB Text algorithm had the highest accuracy. The J48 algorithm gave better results for non-elongated SMS.

Our second experiments compared the different datasets presented in Table 9: ”elongated SMS” and ”non-elongated SMS” (see Tables 6 and 7), ”elongated SMS Dico” corpus and ”non-elongated SMS Dico” corpus (see Section 3.4), ”shortened SMS” corpus for which we removed the repetition of characters of words with an elongation. In this context, for example the elongated word ”Merciii” present in the elongated SMS file became ”Merci” in the ”shortened SMS” file.

	SMO	J48
elongated SMS	46.38	41.77
non-elongated SMS	59.56	63.38
elongated SMS Dico	50.65	46.38
non-elongated SMS Dico	**64.48**	**64.48**
shortened SMS	45.39	41.77

Table 9: Results in terms of accuracy.

[3] Algorithms were applied with the Weka default parameters, for example the polynomial kernel for SMO, the decision tree J48 method, while the Bayesian classification was used as a probabilistic learning method (DMNB Text, and Naive Bayes).

Table 9 shows that the best accuracy value was obtained with the integration of semantic information for the "non-elongated SMS Dico" corpus.

The application of a "shortened" process did not improve the results. Overall, this study highlighted that the integration of semantic knowledge always improves classification.

5. Conclusion

This paper presents an opinion mining approach adapted to SMS. A specific weighting is proposed for features according to their lexical character (presence of an elongation phenomenon) and/or their semantic specificity (presence of the element in a dedicated dictionary).

We plan to use other algorithms in future studies, while also applying other statistical weights to represent textual data (e.g. TF-IDF, OKAPI).

6. Acknowledgements

This study was supported by the SONGES Project[4] (FEDER and Languedoc-Roussillon). Our work is the continuation of the sud4science LR project (http://sud4science.org/) led by Rachel Panckhurst and supported by CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium), MSH-M (Maison des Sciences de l'Homme de Montpellier), and the DGLFLF (Délégation générale à la langue française et aux langues de France).

7. Bibliographical References

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

Amir, S., Almeida, M. B., Martins, B., Filgueiras, J. a., and Silva, M. J. (2014). TUGAS: Exploiting unlabelled data for twitter sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 673–677.

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

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

Cougnon, L. A. and Thomas, F. (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)*, pages 9–11.

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*.

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

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

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

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

Hangya, V., Berend, G., Varga, I., and Farkas, R. (2014). SZTE-NLP: Aspect level opinion mining exploiting syntactic cues. In *Proceedings of 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 610–614. Association for Computational Linguistics and Dublin City University.

Kobus, C., Yvon, F., and Damnati, G. (2008). Normalizing sms: Are two metaphors better than one?. In *Proceedings of the 22Nd International Conference on Computational Linguistics Volume 1*, pages 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. and Turney, P. D. (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*, pages 26–34. Association for Computational Linguistics.

Panckhurst, R., Détrie, C., Lopez, C., C.Moïse, M.Roche, and 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*, pages 107–138.

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

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560. ACM.

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

[4] http://textmining.biz/Projects/Songes