LSIS at SemEval-2017 Task 4: Using Adapted Sentiment Similarity Seed Words For English and Arabic Tweet Polarity Classification

Amal Htait*,** **Sébastien Fournier***,** **Patrice Bellot***,**

* Aix Marseille University
CNRS, ENSAM, Toulon University
LSIS UMR 7296,13397, Marseille, France.

** Aix-Marseille University
CNRS, CLEO OpenEdition
UMS 3287, 13451, Marseille, France.

{amal.htait, sebastien.fournier, patrice.bellot}@openedition.org

Abstract

We present, in this paper, our contribution in SemEval2017 task 4 : ”Sentiment Analysis in Twitter”, subtask A: ”Message Polarity Classification”, for English and Arabic languages. Our system is based on a list of sentiment seed words adapted for tweets. The sentiment relations between seed words and other terms are captured by cosine similarity between the word embedding representations (word2vec). These seed words are extracted from datasets of annotated tweets available online. Our tests, using these seed words, show significant improvement in results compared to the use of Turney and Littman's (2003) seed words, on polarity classification of tweet messages.

1 Introduction

Sentiment Analysis aims to obtain feelings expressed as positive, negative, neutral, or even expressed with different strength or intensity levels. One of the well known extracting sentiment approaches is the lexicon-based approach. A sentiment lexicon is a list of words and phrases, such as *excellent*, *awful* and *not bad*, each is being assigned with a positive or negative score reflecting its sentiment polarity. Therefore, sentiment lexicon provides rich sentiment information and forms the foundation of many sentiment analysis systems (Liu, 2012).

Our system is based on one of the most significant sentiment lexicon classification methods introduced by Turney and Littman (2003). The method is inspired by the semantic similarity measuring and applied to the sentiment analysis field as a sentiment similarity measuring. In a similar method, Kanayama and Nasukawa (2006) worked on detecting a word's sentiment polarity by measuring the difference between its sentiment similarity with a positive seed word and a negative seed word, respectively. This method achieves better results with larger corpora, where there are more chances to find the word (to be classified) near the positive and negative seed words.

In SemEval2017 task 4, we're working with tweets which will lead to deal with slang words and informal phrases. Therefore, the classic seed words suggested by Turney and Littman (2003), listed below in Table 1, will not be very suitable. For example, the word *Superior* is rarely used in the modern ”social media” English, and it is barely found in tweets compared to other seed words. In the tweets dataset of sentiment140 (Go et al., 2009), the word *Superior* is used 42 times, but the word *Nice* is used 23563 times. Thus, for the English tweets polarity classification task, we use the adapted for tweets seed words extracted in our previous work (Htait et al., 2017). And for the Arabic tweets polarity classification task, we apply the same method as in (Htait et al., 2017) to extract Arabic seed words adapted for tweets, to be used in our system.

positive	**negative**
good, nice, excellent, positive, fortunate, correct, superior.	bad, nasty, poor, negative unfortunate, wrong, inferior.

Table 1: The classic seed words suggested by Turney and Littman (2003).

2 Related Work

The use of seed words was the base of many sentiment analysis experiments, some used the concept with supervised or semi-supervised methods.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 718–722,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

For example, Ju et al. (2012) worked on a semi-supervised method for sentiment classification that aims to train a classifier with a small number of labeled data (called seed data). Some other experiments used the concept with unsupervied methods which reduces the need of annotated training data. For example Turney (2002; 2003), which used statistical measures to calculate the similarities between words and a list of 14 seed words (Table 1), such as point wise mutual information (PMI). But we should note that Turney's seed words were manually selected based on restaurant reviews, which have different nature than tweets. Also we find that Maas et al. (2011) used the concept as "bag of words" but with cosine similarity measure on word embedding.

Our previous work (Htait et al., 2017) was on sentiment intensity prediction of tweets segments using SemEval2016 Task7[1] data. We extracted new seed words as more adapted for tweets seed words. We retrieved the most frequent words in Sentiment140 (Go et al., 2009) and then manually filtered the list to eliminate the neutral words. Our tests in (Htait et al., 2017) showed the efficiency of the new seed words over Turney's 14 seed words. Also, they showed that using cosine similarity measure of word embedding representations (word2vec) yields better results than using statistical measures like PMI to calculate the similarities between words. Therefore, and based on the above experiments, we decide to use for our system cosine similarity measure of word embedding representations, but also to use the adapted for tweets seed words from (Htait et al., 2017).

Even though the Arabic language processing faces more challenges than the English language, since words can have transitional meanings depending on position within a sentence and the type of sentence (verbal or nominal) (Farra et al., 2010), we can still find some interesting experiments in lexical-based sentiment analysis: El-Beltagy and Ali (2013) built a sentiment lexicon based on a manually constructed seed sentiment lexicon of 380 words. Using this lexicon, with assigned sentiment intensity score for each value, they were able to calculate the sentiment orientation for a set of tweets in Arabic language (Egyptian dialect). Another paper by Eskander and Rambow (2015) presented a large list of sentiment lexicon for Arabic language called SLSA where each value is associated with a sentiment intensity score. The scores were assigned due to a link created between the English annotation of each Arabic entry to a synset from SentiWordNet (Cambria et al., 2010). For our system in Arabic language, we are following the same method as the system in English language. But since there is no previously created list of adapted for tweets seed words, we create the list following the same method in (Htait et al., 2017), and then use it with cosine similarity measure of word embedding representations.

3 Adapted seed words

3.1 English seed words

In (Htait et al., 2017), seed words were extracted from Sentiment140 dataset (Go et al., 2009). For the positive seeds, a list of the most frequent words in Sentiment140 positive tweets is retrieved and then manually filtered to eliminate the neutral words, and the same is applied for negative seeds. The list of English seed words adapted to tweets is as shown in Table 2.

Positive	Negative
love, like, good, win , lol, hope, best, thanks, funny, haha, god, amazing, fun,beautiful, nice, cute, cool, perfect, awesome, okay, special, hopefully, glad, congrats, excellent, dreams, sunshine, hehe, positive,fantastic, dance, correct, fabulous, superior, fortunate, relaxing, happy,great, kind, laugh, haven, wonderful, yay, enjoying, sweet,	ill, fucking, shit, fuck, hate, bad, break, sucks, cry, damn, sad, stupid, dead, pain, sick, wtf, lost, worst, fail, bored, scared, hurts, afraid, upset, broken, died, stuck, boring, horrible, negative, unfortunate, inferior, unfortunately,poor, need, suck, wrong, evil, missed, sore, alone, crap, hell, tired, nasty.

Table 2: The Tweets Adapted English seed words (Htait et al., 2017).

3.2 Arabic seed words

The Arabic language's experiences, in lexical-based sentiment analysis, were mostly oriented to sentiment lexicons than to seed words. Large lists of sentiment lexicons were built and used for sentiment analysis. For our system, we create a list of seed words following almost the same method as in (Htait et al., 2017). We search for the most common words in positive tweets and in negative tweets from two annotated corpora of Ara-

[1] http://alt.qcri.org/semeval2016/task7/

bic tweets (Arabic Sentiment Tweets Dataset[2] and Twitter data-set for Arabic Sentiment Analysis[3]). Then, to filter the list and to eliminate the neutral words, we use Mohammad et al.'s (2016) list. That list contain 240 positive and negative words of modern standard Arabic, therefore and due to Arabic dialects variety, using that list to filter will create a list of seed words in modern standard Arabic but adapted for tweets, and it can be used independently of dialects. The list of Arabic seed words adapted for tweets is as shown in Table 3.

Positive	Translation	Negative	Translation
خير	benevolent	بال	worn
الجمال	fairness	بشع	ugly
كبير	grand	وسخ	filthy
أعلى	superior	جائر	unjust
حسن	well	عيب	flaw
عظيم	great	خطير	dangerous
رائع	wonderful	حقير	despicable
نادر	exceptive	بايخ	vapid
جمال	beauty	حزين	sad
كريم	generous	قذر	dirty
أعظم	greatest	هائل	massive
نبيل	noble	مقرف	nasty
جميل	beautiful	باطل	invalid
صالح	valid	تافه	trifle
دقيق	accurate	ملعون	damned
مشرق	bright	مرفوض	unacceptable
طيب	delicious	مسكين	poor
حلو	sweet	فاسد	corrupt
جيد	good	مؤسف	regrettable
عبقري	genius	فظيع	horrible

Table 3: The Tweets Adapted Arabic seed words.

4 System of Sentiment classification

Our System is based on sentiment similarity cosine measure with Word Embedding representations (word2vec). For the English language, we use twitter word2vec model by Godin et al (Godin et al., 2015), since best results were achieved using that model in sentiment intensity prediction with the adapted seed words (Htait et al., 2017). This model is a word2vec model trained on 400 millions tweets in English language and it has word representations of dimensionality 400. For the Arabic language, there is no twitter word2vec model available online (to the best of our knowledge). Therefore, we collect 42 millions tweets in Arabic language from archived twitter streams[4] to create our twitter word2vec model.

In Figure 1, we have the work flow of our system for tweets sentiment classification. First, each tweet is cleaned by removing links, user names, stop words, numeric tokens and characters except the common emoticons: ”:-)”, ”:-(”, ”:)”, ”:(”, ”:’(”. Also, words with repetitive characters are replaced by the corrected ones (e.g. coooool by cool). After that, the tweet is segmented into tokens or words. The similarity between each word with positive seed words and negative seed words is calculated using gensim tool[5] with the previously mentioned word2vec models for both languages English and Arabic.

Having the sentiment score of each word in a tweet, we aggregate by sum to combine these values. The final score specify the tweet's polarity. After many tests on old SemEval data (task ”Message Polarity Classification” of 2013 and 2014), we found that the best scores achieved are by considering the following: if the score is higher than 1, the tweet is considered positive, else if the score is lower than -2, the tweet is considered negative, else it is considered neutral.

To test the efficiency of the adapted seed words on tweets polarity classification, we apply our system on SemEval data for the task : ”Sentiment Analysis in Twitter” of years 2013[6] and 2014[7], using Turney's seed words (in Table1), and the adapted seed words. The Table 4, along with the results, shows clearly how the use of the adapted for tweets seed words increase the results compared to Turney's seed words.

2013	**AvgF1**	**AvgR**	**Acc**
Turney	0.262	0.381	0.480
Adapted	**0.564**	**0.571**	**0.508**
2014	**AvgF1**	**AvgR**	**Acc**
Turney	0.303	0.383	0.511
Adapted	**0.589**	**0.553**	**0.552**

Table 4: The comparison between Turney's seed words and the adapted seed words on semEval task's data of years 2013 and 2014.

The results of our participation at SemEval2017 Task4 (subtask A) for English and Arabic lan-

[2] http://www.mohamedaly.info/datasets/astd

[3] https://archive.ics.uci.edu

[4] https://archive.org/details/twitterstream

[5] https://pypi.python.org/pypi/gensim

[6] https://www.cs.york.ac.uk/semeval-2013/task2.html

[7] http://alt.qcri.org/semeval2014/task9/

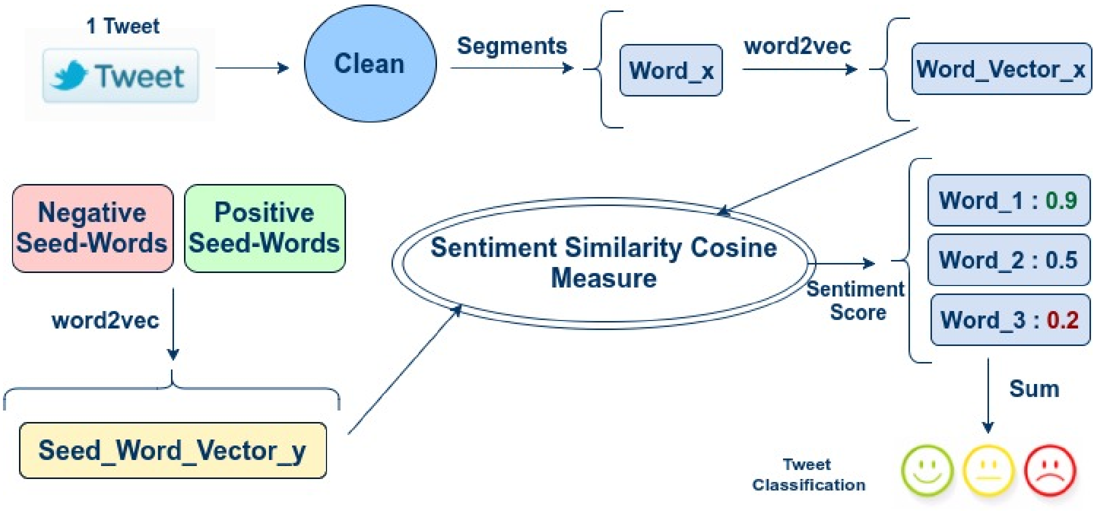

Figure 1: The work flow of tweets sentiment classification.

guages are in Table 5, with the best results accomplished in the subtask A.

English	Team	AvgF1	AvgR	Acc
	BB_twtr LSIS	0.685 0.561	0.681 0.571	0.658 0.521
Arabic	**Team**	**AvgF1**	**AvgR**	**Acc**
	NileTMRG LSIS	0.610 0.469	0.583 0.438	0.581 0.445

Table 5: The results at semEval2017 Task 4 subtask A - for English and Arabic Languages.

5 Conclusion

In this paper, we present our contribution in SemEval2017 task4: Sentiment Analysis in Twitter, subtask A: Message Polarity Classification, for English and Arabic languages. Our system is based on a list of sentiment seed words adapted for tweets, used in sentiment similarity cosine measure with word embedding representations (word2vec). Although the results are encouraging, further investigation is required concerning the detection of negations (e.g. not) and intensifiers(e.g. very) in the tweets, due to their big effect on reversing the polarity of a tweet.

Acknowledgments

This work was supported by the French program Investissements d'Avenir Equipex ”A digital library for open humanities” of OpenEdition.org.

References

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*. volume 10.

Samhaa R El-Beltagy and Ahmed Ali. 2013. Open issues in the sentiment analysis of arabic social media: A case study. In *Innovations in information technology (iit), 2013 9th international conference on*. IEEE, pages 215–220.

Ramy Eskander and Owen Rambow. 2015. Slsa: A sentiment lexicon for standard arabic. In *EMNLP*. pages 2545–2550.

Noura Farra, Elie Challita, Rawad Abou Assi, and Hazem Hajj. 2010. Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, pages 1114–1119.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1(12).

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations. *ACL-IJCNLP* 2015:146–153.

Amal Htait, Sébastien Fournier, and Patrice Bellot. 2017. Identification automatique de mots-germes pour l'analyse de sentiments et son intensité. In *RJCRI*. Marseille, France.

Shengfeng Ju, Shoushan Li, Yan Su, Guodong Zhou, Yu Hong, and Xiaojun Li. 2012. Dual word and document seed selection for semi-supervised sentiment classification. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, pages 2295–2298.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 355–363.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 142–150.

Saif Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. Sentiment lexicons for arabic social media. In *LREC*. Portoro, Slovenia.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 417–424.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21(4):315–346.