Creating multi-scripts sentiment analysis lexicons for Algerian, Moroccan and Tunisian dialects

Karima Abidi, Kamel Smaïli

Creating multi-scripts sentiment analysis lexicons for Algerian, Moroccan and Tunisian dialects

K. Abidi and K. Smaïli

Loria - University Lorraine, France
{kabidi, ksmaili}@loria.fr

Abstract. In this article, we tackle the issue of sentiment analysis in three Maghrebi dialects used in social networks. More precisely, we are interested by analysing sentiments in Algerian, Moroccan and Tunisian corpora. To do this, we built automatically three lexicons of sentiments, one for each dialect. Each lexicon is composed of words with their polarities, a dialect word could be written in Arabic or in Latin scripts. These lexicons may include French or English words as well as words in Arabic dialect and standard Arabic. The semantic orientation of a word represented by an embedding vector is determined automatically by calculating its distance with several embedding seed words. The embedding vectors are trained on three large corpora collected from YouTube. The proposed approach is evaluated by using few existing annotated corpora in Tunisian and Moroccan dialects. For the Algerian dialect, in addition to a small corpus we found in the literature, we collected and annotated one composed of 10k comments extracted from Youtube. This corpus represents a valuable resource which is proposed for free [1].

Keywords: Maghrebi dialect · Word embedding · Semantic orientation.

1 Introduction

To understand the requirements of users, clients or people in general, it is necessary to mine social media [3, 4, 30] and to develop automatic tools allowing a systematic analysis of the contents. One can then extract useful information that could be used in marketing advising, political views, movies reviews, etc. Henceforth, proposing methods to understand opinions is necessary and it is considered as a challenging issue especially for under-resourced languages such as Maghrebi dialects.
In this article, we will address the issue of developing a method allowing to analyse sentiments in the three following Maghrebi dialects: Algerian, Moroccan and Tunisian. The problem is that these dialects are under-resourced because they are not formal and not official languages. Basically, Arabic dialects are founded on Modern Standard Arabic (MSA), but not only. The originality of this work

[1] https://smart.loria.fr/corpora

is to propose a sentiment analysis tackling two issues frequent in Algerian, Moroccan and Tunisian dialects: the code-switching nature of a document and its multi-script form.
In order to explain the importance of this research work, let's give in the following the different particularities of the Maghrebi dialects.

The origin of Maghrebi dialects is mainly Standard Arabic, but not only. For practical reasons, several morpho-syntactic rules of MSA are not respected in Arabic dialects. This means that it is difficult to use the amount existing NLP resources developed for MSA to process Arabic dialects.

The vocabularies of the Arabic dialects evolve continuously by introducing new words, that could be considered as gibberish such as the word *papicha* that means *beautiful girl* in Algerian dialect. And as in any other language, Maghrebi dialects can borrow new words and adapt them phonologically to the local dialect such as: كوزِينة (borrowed from the French word *cuisine* that means *kitchen*).

Another particularity of Maghrebi dialects is the fact that people can write Arabic by using multiple scripts: Arabic and Latin [6, 7]. In addition, in social media people can use digits when they write in Latin script to represent sounds that do not exist in French or English, such as ع which is replaced by the digit 3.

In addition to these phenomena, in north Africa, code-switching is common in conversations. One can mix in the same sentence local Arabic, MSA, and foreign languages, such as French or English. In the following, we give an example:

تبارك الله عليك *merci pour la vidéo* عجبني *les boucles d'oreilles* منين خديتهم **thanks.**

Translation : "*God bless you, thank you for this video, I really liked the earrings where did you buy them. Thank you.*"

In this sentence, one switched from Moroccan dialect written in Arabic script, to French, then to Arabic, then again to French then once again to Arabic and finally to English.

Only few works addressing the issue of sentiment analysis in Maghrebi dialects do exist. But most of them have ignored the problems already cited and have concentrated on sentiment analysis in Maghrebi texts written only in Arabic script. In this article, we propose a method that allows to create automatically sentiment lexicons for Arabic dialects taking into account all the phenomena aspects related to Maghrebi dialects. This approach could be adapted to any Arabic dialect and also to any other low-resourced languages.

The rest of the paper is organized as follows: Section 2 is dedicated to the related work, while Section 3 examines the corpus we harvested. In Section 4, we discuss the proposed method to analyse sentiment of Maghrebi dialect. In Section 5, we present the different used corpora and the experimental results and finally we conclude.

2 Related work

Many studies have been conducted to address the issue of sentiment analysis in Arabic documents [2, 22, 8]. Researchers proposed various interesting approaches, that we can classify into two categories: machine learning techniques [10, 20, 12, 28, 18] and lexicon-based approaches [27, 15, 1, 14, 21]. Unfortunately, most of these methods are not directly reusable for Arabic dialects for the reasons mentioned in the introduction. In this section, we discuss the research works proposed to analyse sentiments in Arabic dialects and we will focusing on those concerning the three Maghrebi ones studied in this article.

In the Arabic dialect sentiment analysis literature, several works have used machine learning techniques to address this issue. These methods require a significant amount of pre-annotated corpora to train a good classifier that is able to distinguish between positive and negative documents.

In [16], the authors proposed a deep learning model based on Long short-term memory (LSTM) architecture to identify the sentiments of documents written in Egyptian and Emirati dialects. To train this model, the authors collected and annotated a corpus of 470k tweets. This model achieved an accuracy of 70% and 63% for Egyptian and Emirati, respectively. The authors of [9] proposed a model that combines LSTM with a convolutional neural network architectures. They used two existing annotated corpora extracted from Twitter to train the proposed model, which is composed of 10k and 2k of tweets written in Egyptian and Levantine, respectively. The method achieved an accuracy of more than 85%. Deep LSTM architecture has been also used by the authors of [23] to tackle the issue of sentiment analysis in Tunisian Dialects. The authors trained the model on a Tunisian corpus composed of 17k and the method achieved an accuracy of 90%.

Unlike the machine learning techniques, in the based-lexicon method, the global sentiment of a document is estimated by calculating the semantic orientation of the words appearing within the text. This approach requires the use of a lexicon of words with their polarities (positive and negative). In this approach, the need of sentiment lexicons is crucial for analysing documents in terms of opinions, that is why the authors in [17] created semi-automatically a sentiment lexicon, where 45% of the entries are Egyptian while 55% are words of Modern Standard Arabic. A sentiment lexicon for the Khaliji dialect has been built manually by exploring and labelling the words of a Saudi dialect twitter corpus (SDTC) [11]. The authors of [24] built a lexicon for Algerian dialect by translating manually an existing Egyptian polarity lexicon. In [13], the authors proposed an approach for emotion analysis of Tunisian comments posted in Facebook by using an emotion dictionary created automatically.
Actually, only a limited number of researches have been carried out for sentiment analysis in the three Maghrebi dialects, while a majority of research in this scope are dedicated to texts written only in Arabic script.

3 The collected dataset

In this work, we are interested by the Algerian, Moroccan and Tunisian dialects. That is why, we extracted three large corpora from YouTube by using the approach proposed in [5]. This method crawls (by using the API[2]) the posts of videos using specific hashtags related to each country.

In Table 1, we give some figures about the collected data, where $|C|$ indicates the number of comments, $|W|$ the number of words and finally $|V|$ the size of the vocabulary. We mention that these statistics concern the data obtained after the cleaning process.

Table 1: Statistics of the harvested corpora.

	Algerian (M)	**Moroccan (M)**	**Tunisian(M)**
$\|C\|$	1.61	1.60	1.26
$\|W\|$	23	22	17
$\|V\|$	1.2	1.3	1

4 The sentiment analysis method

In this work, we propose a lexicon-based approach to analyse the sentiments of Maghrebi comments extracted from social networks. In this approach, the polarity of a text can be obtained on the ground of the polarity of the words that compose it. To do this, a lexicon of words, where each entry is associated to its polarity is necessary.
Because, in the Maghrebi dialects people use Latin and Arabic scripts and foreign languages to post their comments, we aim, in this work to handle this issue by building a multi-script and multilingual sentiment lexicon in order to analyze the sentiments of the collected corpus. A word and its polarity constitute an entry in the lexicon. Each word of this lexicon can be written in Arabic or Latin script and it can belong to one of the following languages: one of the three Maghrebi dialects, MSA, French or English as in the table 2. Concerning the polarity of each entry, we determined it automatically by using an approach similar to the one used in [29]. In this method, the authors proposed to tag the words by using the polarities of a small list of words called *seed words* for which the polarity is assigned by hand. A word is attached to the dominant polarity of the closest seed words. For example, a word is considered positive if it is closer, in terms of distance, to positive seed words than to negative words.

In the following, we will detail the two main steps of the method we used to assign a polarity to an entry of a sentiment lexicon.

[2] https://developers.google.com/YouTube

Table 2: Few examples concerning the different forms of a word

Script	language	Word	Meaning
Latin	Algerian dialect	Chaba	Beautiful
Arabic	Algerian dialect	شَابة	Beautiful
Arabic	MSA	كَاذب	Liar
Latin	MSA	Kadib	Liar
Latin	English	Like	Like
Arabic	English	لَايك	Like

4.1 Seed words identification

In [29], the authors used a list of seed words made up of seven positive words and seven negative words. In our case, for each dialect, we manually annotated a list of forty seed words for each of the polarities. The seed words have been selected, from a list of the most frequent not neutral words of the collected corpora. Then, we assigned manually to each of them its corresponding polarity. The number of seed words retained is relatively high compared to the experiment carried out by the authors of [29]. This is due to the fact that we want to cover as many words as possible since we deal with multi-script and multilingual words in our corpora. In the table 3 we give some examples of these retained seed words.

Table 3: Some examples of positive and negative seed words for the three dialects.

Algerian		Moroccan		Tunisian	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
chaba (*Pretty*)	شيَات (*Groveller*)	روعة (*Wonderful*)	Problème (*Problem*)	To9tool (*so beautiful*)	مَاسط (*Boring*)
Bravo	Mosakh (*Disgusting*)	Hbiba (*Sweetie*)	ينعل (*Cursed*)	Ma7lek (*How beautiful you are*)	جَاهل (*Ignorant*)
هَايل (*Excellent*)	Na3ja (*A weak personality*)	Tbarklah (*Marvelous*)	Himar (*Donkey*)	تهبل (*Wonderful*)	حيوَان (*Beast*)
الصحة (*Health*)	Roukhs (*Asshole*)	كنحمَاق (*I love*)	حرَام (*Not good*)	حلوه (*Delicious*)	حنش (*Snake*)

4.2 Estimation of the polarity of words

We propose to calculate the semantic orientation of a word w according to the difference between its closest positive seed words SW_{pos} and its closest negative seed words SW_{neg}. We estimate the degree of closeness between two words one of which is a seed word by using the cosine similarity as in the formula we propose in 1:

$$SO(w) = \sum_{w_p \in SW_{pos}} Cos(w, w_p) - \sum_{w_n \in SW_{neg}} Cos(w, w_n) \quad (1)$$

w is considered as positive if $SO(w)$ is positive, similarly w is considered as negative whether its orientation is negative. The similarity of the word w and a seed word is estimated by using the cosine measure between their corresponding embedding vectors. The embedding vectors are generated by using the Continuous Bag-of-Words (CBOW), one of the method of the Word2Vec approach proposed by Mikolov [26]. The training has been achieved on the three large corpora presented in section 3. This led to the creation of a lexicon of sentiments for each of the dialects cited in this research. In the table 4, we detail each of them.

Table 4: Some figures about the lexicons of sentiments

	Algerian	**Moroccan**	**Tunisian**
Number of entries	11243	23405	10810
Positive words	8372	2326	19128
Negative words	2871	8484	4277

In table 5, we give some examples extracted from our lexicons.

Table 5: Some examples of positive and negative words extracted from the three lexicon

Word	**Translation**	**Polarity**	**Lexicon**
ضحكاتنى	She makes me laughing	Positive	Moroccan
Kanbghiwk	We like you	Positive	Moroccan
البرهوش	Stubborn	Negative	Moroccan
frahnalk	Happy for you	Positive	Algerian
حركي	Groveller	Negative	Algerian
wetek	It suits you very well	Positive	Tunisian
ثعالب	Crafty	Negative	Tunisian

5 Experimentation

We used the created lexicons to evaluate the polarity of the four following labeled corpora.

– **ElecMorocco**: is a Moroccan corpus extracted from Facebook and annotated by the authors of [19]. The main topic of this corpus concerns the local elections. It is constituted by 6389 positive and 4367 negative comments. This corpus contains only comments written in Arabic script.

– **TSAC**: is a Tunisian corpus collected from comments posted on official Facebook pages of Tunisian radios and TV channels [25]. It is composed of 5081 positive and 6514 negative comments written in Arabic and Latin scripts.

– **CorpusAlg**: is an Algerian corpus extracted from different Facebook pages, it contains 5079 posts, among them 3032 are positive comments [24]. The comments in this corpus are written in Latin and Arabic characters.

– **SentAlg**: The previous Algerian corpus (CorpusAlg) is small in comparison to the two others, that is why we decided to collect and annotate manually 10k of comments extracted from Algerian YouTube comments. This achieved a corpus of positive and negative comments of 5562 and 4438 respectively. All of the comments are written in Arabic and Latin script.

Table 6 summarizes the figures of the different corpora used in our experimentation.

Table 6: Some figures about the different corpora.

	ElecMorocco	**TSAC**	**CorpusAlg**	**SentAlg**
Positive comments	3523	5081	3032	5562
Negative comments	6431	6514	2047	4438

5.1 Results

As mentioned before, we used in this work a sentiment analysis method which is based on sentiment lexicons and annotated corpora. In this method, the aim is to identify in the sentence under analysis, the words that exist in the lexicon and to take into account the corresponding polarities in the opinion to assign to the sentence. Then the evaluation is estimated by using scores such as: accuracy, recall, precision, F-measure, etc. In table 7, we give the achieved results in terms of Recall and Precision on the corpora listed above.

Table 7: Experimental results on the three Maghrebi corpus.

	Recall (%)	**Precision(%)**	**F-measure (%)**
ElecMorocco	59.29	63.23	61.19
TSAC	64.03	63.78	63.90
CorpusAlg	67.92	67.15	67.53
SentAlg	**79.78**	**80.79**	**80.28**

The results show that the weakest performance concerns the Moroccan corpus. An analysis of this corpus shown that this later contains a lot of sentences in Modern Standard Arabic, while our training was done on a corpus extracted from the comments of Youtube posted by Moroccan mostly in their dialect [6]. Another explanation of these results is due to the fact that ElectMoroccan contains only comments written in Arabic script which is not the case of the training corpus. Concerning the Tunisian and the Algerian test corpora TSAC and CorpusAlg, respectively, thematically, they are far from the crawled corpora used for the training. That is why the performances are reasonable, but they are not as good as the results achieved on the corpus SentAlg. We recall, that we used a lexicon-based approach for sentiment analysis, but unfortunately we have not found any available sentiment lexicon for the three studied dialects, that is why we created them automatically. In the opposite, SentAlg is a corpus which is similar thematically to the training corpus, which explains why the performances are higher than those obtained for the others.

6 Conclusion

We proposed, in this article, a lexicon-based approach to analyse sentiments of three Maghrebi dialects, namely Algerian, Moroccan and Tunisian. The dialects lexicons used to classify the documents in terms of sentiments were created automatically. The approach, we proposed depends on a predefined list of 80 polarity seed words for each dialect, selected manually. Then, using a similarity measure, we estimated the proximity between an embedding word and the list of the embedding seed words.

One of the originality of this method is that it allowed to create multi-script and a multi-lingual sentiment lexicons for Algerian, Moroccan and Tunisian dialects which contain 11.2k, 23.4k and 10.8k entries, respectively. These sentiment lexicons were used to classify, in terms of polarities, three test datasets. This approach has been also tested on an Algerian corpus we collected and labelled manually. The performances we achieved depend on the quality of the corpora and they vary between 61.99 and 80.28 in terms of F-measure.

References

1. Abd-Elhamid, L., Elzanfaly, D., Eldin, A.S.: Feature-based sentiment analysis in online arabic reviews. In: 2016 11th International Conference on Computer Engineering Systems (ICCES) (2016)
2. Abdul-Mageed, M., Diab, M., Korayem, M.: Subjectivity and sentiment analysis of modern standard Arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011)
3. Abdul-Mageed, M., Diab, M.T., Kübler, S.: SAMAR: subjectivity and sentiment analysis for arabic social media. vol. 28, pp. 20–37 (2014)

4. Abidi, K., Fohr, D., Jouvet, D., Langlois, D., Mella, O., Smaïli, K.: A Fine-grained Multilingual Analysis Based on the Appraisal Theory: Application to Arabic and English Videos. vol. Communications in Computer and Information Science book series (CCIS, volume 1108), pp. 49–61. Springer, Nancy, France (2019)
5. Abidi, K., Menacer, M.A., Smaili, K.: CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube. In: 18th Annual Conference of the International Communication Association (Interspeech). Conference of the International Communication Association (Interspeech), Stockholm, Sweden (2017)
6. Abidi, K., Smaïli, K.: An empirical study of the Algerian dialect of Social network. In: ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing. Casablanca, Morocco (2017), https://hal.inria.fr/hal-01659997
7. Abidi, K., Smaïli, K.: An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings. In: 11th edition of the Language Resources and Evaluation Conference, LREC 2018. Miyazaki, Japan (2018), https://hal.archives-ouvertes.fr/hal-01718110
8. Abo, M.E.M., Raj, R.G., Qazi, A.: A review on arabic sentiment analysis: State-of-the-art, taxonomy and open research challenges (2019)
9. Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., Dobnik, S.: Lstm-cnn deep learning model for sentiment analysis of dialectal arabic. In: Smaïli, K. (ed.) Arabic Language Processing: From Theory to Practice (2019)
10. Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El Hajj, W., Bashir Shaban, K.: Deep learning models for sentiment analysis in Arabic. In: Proceedings of the Second Workshop on Arabic Natural Language Processing (Jul 2015)
11. Al-Thubaity, A., Alqahtani, Q., Aljandal, A.: Sentiment lexicon for sentiment analysis of saudi dialect tweets. vol. 142, pp. 301–307 (2018), arabic Computational Linguistics
12. Ali, A.E., Stratmann, T.C., Park, S., Schöning, J., Heuten, W., Boll, S.: Measuring, understanding, and classifying news media sympathy on twitter after crisis events. vol. abs/1801.05802 (2018)
13. Ameur, H., Jamoussi, S., Ben Hamadou, A.: Exploiting emoticons to generate emotional dictionaries from facebook pages. In: Czarnowski, I., Caballero, A.M., Howlett, R.J., Jain, L.C. (eds.) Intelligent Decision Technologies 2016 (2016)
14. Awwad, H., Alpkocak, A.: Performance comparison of different lexicons for sentiment analysis in arabic. In: 2016 Third European Network Intelligence Conference (ENIC) (2016)
15. Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W.: A large scale arabic sentiment lexicon for arabic opinion mining. In: ANLP@EMNLP (2014)
16. Baly, R., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., Shaban, K.B., El-Hajj, W.: Comparative evaluation of sentiment analysis methods across arabic dialects. Procedia Computer Science **117**, 266–273 (2017), arabic Computational Linguistics
17. El-Beltagy, S.: Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic (05 2016)
18. Elfaik, H., Nfaoui, E.H.: Deep bidirectional lstm network learning-based sentiment analysis for arabic text. vol. 30, pp. 395–412 (2021)
19. Elouardighi, A., Maghfour, M., Hammia, H., Aazi, F.Z.: Analyse des sentiments à partir des commentaires facebook publiés en arabe standard ou dialectal marocain par une approche d'apprentissage automatique. In: Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018. pp. 329–334 (2018)
20. Heikal, M., Torki, M., El-Makky, N.: Sentiment analysis of arabic tweets using deep learning. vol. 142, pp. 114–122 (2018), arabic Computational Linguistics

21. Htait, A., Fournier, S., Bellot, P.: Identification semi-automatique de mots-germes pour l'analyse de sentiments et son intensité. In: COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings. pp. 415–424 (2017)
22. Ibrahim, H.S., Abdou, S., Gheith, M.: Sentiment analysis for modern standard arabic and colloquial. vol. abs/1505.03105 (2015)
23. Jerbi, M.A., Achour, H., Souissi, E.: Sentiment analysis of code-switched tunisian dialect: Exploring rnn-based techniques. In: Arabic Language Processing: From Theory to Practice - 7th International Conference, ICALP 2019, Nancy, France, October 16-17, 2019, Proceedings. pp. 122–131 (2019)
24. Mataoui, M., Zelmati, O., Boumechache, M.: A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. vol. 110, pp. 55–70 (2016)
25. Medhaffar, S., Bougares, F., Estève, Y., Belguith, L.H.: Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In: Proceedings of the Third Arabic Natural Language Processing Workshop, WANLP 2017@EACL, Valencia, Spain, April 3, 2017. pp. 55–61 (2017)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (Workshop) (2013)
27. Mohammad, S., Dunne, C., Dorr, B.J.: Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: EMNLP (2009)
28. Nejjari, M., Meziane, A.: Sahar-lstm: An enhanced model for sentiment analysis of hotels'arabic reviews based on lstm (2020)
29. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association (2003)
30. Yimam, S.M., Alemayehu, H.M., Ayele, A., Biemann, C.: Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In: Proceedings of the 28th International Conference on Computational Linguistics (Dec 2020)