Sentiment Analysis of Code-Switched Tunisian Dialect: Exploring RNN-Based Techniques

Mohamed Amine Jerbi[1](✉), Hadhemi Achour[1], and Emna Souissi[2]

[1] Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000 Le Bardo, Tunisia
mohamedaminejerbil@gmail.com, Hadhemi_Achour@yahoo.fr
[2] Université de Tunis, ENSIT, 1008 Montfleury, Tunisia
emna.souissi@ensit.rnu.tn

Abstract. With the increasing use of social networks and the multilingualism that characterizes the Internet in general and the social media in particular, an increasing number of recent research works on Sentiment Analysis and Opinion Mining are tackling the analysis of informal textual content, which includes language alternation, known as code-switching. To date, very little work has addressed in particular, the analysis social media of the Tunisian dialect, which is characterized both by a frequent occurring of code-switching and by a double script (Arabic and Latin) when written on the social media. Our study aims to explore and compare various classification models based on RNNs (Recurrent Neural Networks), precisely on LSTM (Long Short-Term Memory) neural networks.

Keywords: Sentiment analysis · Code-switching · Tunisian dialect · Social media · Deep learning · RNN · LSTM · Bi-LSTM · Deep-LSTM

1 Introduction

Code-switching (CS) is the alternation of at least two linguistic codes in a single conversation. It is defined in [1] as «*the mixing, by bilinguals (or multilinguals), of two or more languages in discourse, often with no change of interlocutor or topic.*». CS may occur at any level of linguistic structure such as a clause, a single sentence, a constituent or even a word.

This phenomenon is very common on social networks, blogs and forums, especially within multilingual communities. In terms of NLP tasks, code-switching data reveals a number of problems requiring dedicated tools handling the mixed data specificities [2]. In particular, when it comes to sentiment analysis within this kind of textual content, emotions and opinions can be expressed in several different ways in different languages, which can complicate the automatic sentiment detection and opinion mining tasks [3].

In this work, we focus on sentiment analysis (SA) in Tunisian dialect textual productions that are generated on the social media and which are characterized by a very frequent use of code-switching. The automatic processing of the Tunisian dialect on social networks is a challenging task and is still in its infancy, especially when

K. Smaïli (Ed.): ICALP 2019, CCIS 1108, pp. 122–131, 2019.
https://doi.org/10.1007/978-3-030-32959-4_9

considering the very low availability of linguistic resources and dedicated NLP tools, despite the growing interest it arouses among many researchers in the last few years [4].

The Tunisian dialect (TD), known as the "*Darija*" or "*Tounsi*", the TD is a variant of the Arabic language, quite different from the MSA, because it derives from different substrates and a mixture of several languages: Punic, Berber, Arabic, Turkish, French, Spanish and Italian [4]. When used on the social media, TD is characterized by an unformal writing style, enclosing emoticons, elongated specific words and a simplified syntax with misspellings. In addition, Tunisian social media users produce a content using a mixture of many languages (French, English, standard Arabic and Tunisian dialect) and they are able to easily switch between them. TD can be written using the Arabic or the Latin script with the use of numeric digits replacing some Arabic letters. The phenomenon of code-switching is mainly a result of multilingualism. In Tunisia, CS is frequent and represents a feature of the local way of speech. In text, CS generally occurs between MSA and TD when it is written in Arabic script and between French/English and TD when it is written with Latin alphabet. Furthermore, some texts (especially in the social media and advertising spots) mix both Arabic and Latin scripts.

Till today, very little work has been done on TD sentiment analysis. Moreover, CS and multiscript writing phenomenon have not always been taken into account in previous works on TD SA. As for the adopted approaches, they were mainly based on classical machine learning classifiers.

In this work, we propose an approach based on deep learning techniques for the SA of code-switched TD data, collected from the social media. Given the lack of available NLP tools for the TD, we propose as a first exploratory study of this problem, to evaluate the performance of a sentiment classification (Positive/Negative), based on Recurrent Neural Networks (RNNs) and word embeddings, and using the freely available corpus TSAC[1] (Tunisian Sentient Analysis Corpus) [5]. TSAC includes a set of opinions extracted from social networks, written in both Latin and Arabic scripts and containing code switching. This will allow us to compare our results with those obtained by Mdhaffar et al. [5] who have experienced Naïve Base (NB), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classifiers using TSAC corpus. For our part, we propose to experiment and compare the classifiers LSTM, bi-LSTM, deep-LSTM and deep-bi-LSTM, which, to our knowledge, have not yet been explored in the case of Tunisian Dialect SA.

The next section of this paper is dedicated to a brief review of previous work carried out on the SA of code-switched texts and a review of work on the Tunisian dialect SA. In Sect. 3, we present our approach for classifying code-switched Tunisian dialect comments. We present in particular, the principle of each proposed classifier. The experiments and results obtained are presented in Sect. 4 and Sect. 5 is dedicated to the conclusion and future perspectives.

[1] TSAC Corpus is available at: https://github.com/fbougares/TSAC.

2 Related Work

This section presents a brief literature review on SA in code switched text. It then, focuses on a comprehensive review of works related to SA of Tunisian dialect textual content, in order to study how the code-switching phenomena, widely present in dialectical writings, have been treated so far.

2.1 Code-Switched Text SA

Several works have tackled SA in code-switched data using a variety of approaches. Mataoui et al. [6] proposed a lexicon-based SA approach for the Algerian Vernacular Arabic. They used three lexicons: Keywords lexicon, negation words lexicon and intensification word Lexicon. Their process is divided into four models: common phrases similarity computation module, pre-processing module, language detection & stemming module, and Polarity computation module. The experimental results showed that their system obtains an accuracy of 79.13%.

Other works have adopted a machine learning-based approach, such as Vilares et al. [7] who compared in their study, different machine learning algorithms to perform multilingual polarity classification in three different environments. Their goal was to compare the performance of supervised models. They trained all their classifiers using a L2 regularized logistic regression. They proposed an English monolingual model, a Spanish monolingual model (using language identification tools), and a multilingual model trained on a mixed dataset that does not need any language recognition step. The latter seems to outperform the monolingual models. In this category of approaches that are based on machine learning methods, we can also cite Barman et al. [8], Supraja and Rao [9], and Vilares et al. [10].

Sentiment analysis of code-switched text has also been tackled using deep learning approaches. Indeed, Wang et al. [3] collected at first their dataset based on Chinese and English mixed code-switching text from Chinese social media. They then explored the challenges of emotions in code-switching text, more precisely, the monolingual and bilingual information in each post. They worked on capturing the informative words from the code-switching context. They dealt with these challenges by proposing a Bilingual Attention Network model to integrate the attention vectors in order to predict the emotion. They first used an LSTM network to build a document representation for each post. Secondly, the document representation is projected into three vectors by collecting the representation of informative words from monolingual and bilingual context. And thirdly a full-connected layer is used to integrate the three attention vectors, and predict the emotion using the Softmax function. Joshi et al. [11], worked on a Hindi-English code-mixed annotated dataset for sentiment analysis, extracted from Facebook popular pages in India. They introduced how to learn sub-word level representations in LSTM (Subword-LSTM) deep neural network architecture instead of word-level and character-level representations. Their Subword-LSTM system using character embeddings as a first layer provides a higher F-score than Char-LSTM and captures more sentiments in Code Mixed and other varieties of noisy data from the social media.

Moreover, it should be noted that, some research works have used both machine learning and deep learning methods. Baly et al. [12], for example used as a corpus, the Arabic Sentiment Twitter Data (ASTD) [13]. They explored the performance of the feature engineering approach using SVM with Word n-grams and Character n-grams and a deep learning approach using Recursive Neural Tensor Network (RNTN) with the sentiment treebank, on Arabic Twitter. The results showed that the RNTN achieves best performance, and confirms the advantage of recursive deep learning, over models that apply feature engineering.

2.2 Tunisian Dialect SA

Ameur et al. [14] highlighted the social network influence on the events of the Tunisian revolution and focused on their work on emotion analysis of Tunisian Facebook pages. They collected comments from Facebook pages in order to analyze sentiments written in Tunisian dialect, and proposed a method for dynamic emotional dictionaries construction based on the use of a language identification tool and the use of emotion symbols (emoticons) as indicators of sentiment polarity, without using external linguistic resources and they considered nine classes: *surprised*, *satisfied*, *happy*, *gleeful*, *romantic*, *disappointed*, *sad*, *angry* and *disgusted*.

Sayadi et al. [15] used a manually annotated dataset extracted from Twitter and compared five different classifiers: Naive Bayes (NB), Support Vector Machines (SVM), k-Nearest Neighbor (NN), Decision Trees (DT) and Random Forest (RF) using the information Gain feature selection method applied to Sentiment Analysis. They tested the trained classifiers on their dataset composed of Tunisian Dialect and Modern Standard Arabic. The dataset was a collection of 10 k tweets written in Arabic letters. SVM classifiers trained with 1-gram, 2-grams and 3-grams as features, gave the best results.

Mdhaffar et al. [5] proposed a freely available and annotated Tunisian Dialect corpus of 17 k comments, extracted from Facebook, called TSAC (Tunisian Sentiment Analysis Corpus). They focused on SA of the Tunisian dialect applying Machine Learning techniques to determine the polarity of comments, such as Support Vector Machines (SVM), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). They trained the Tunisian dialect SA system and obtained an accuracy of 77% with SVM, 78% with MLP and 58% with NB.

For their part, Mulki et al. [16] have proposed to study the impact of several preprocessing techniques on TD SA. They experimented two SA models: a supervised machine learning based model (SVM and NB) and a lexicon-based model. They showed the importance of the stemming, emoji and negation tagging preprocessing tasks for the TD SA. They also showed that adding name entities tagging to these tasks leads to best results.

2.3 Discussion

From this brief state of the art, we can notice that several approaches have been proposed to deal with the sentiment analysis of code-switched text, ranging from linguistic approaches, machine learning, deep learning and hybrid approaches. It seems that approaches integrating deep learning methods are leading to the best performance. We can also deduct that, the preferred approach in several studies, which seems to be the most effective, is that which considers the multilingual text in its entirety, without going through language detection and the use of monolingual classifier for each language detected in the text.

As for the Tunisian Dialect SA which is our goal in this work, we can notice that there is very little literature that addresses this problem. Among the four surveyed works, three of them (Mdhaffar et al. [5], Sayadi et al. [15] and Mulki et al. [16]) have had as objective, to classify TD comments by sentiment. We however notice, that only, Mdhaffar et al. [5] and Mulki et al. [16] considered a code-switched corpus enclosing both Arabic and Latin scripts, while Sayadi et al. [15] have only dealt with the Arabic script, and haven't considered the Latin script which is massively used in the TD writings nor the phenomena of code-switching which is also very frequent on the social media. As for the adopted approaches, we can see that deep learning-based methods have been practically not explored with this language and TD SA has been approached, mainly by experiencing classical machine learning classifiers. For this reason, we propose in this work to study and evaluate some deep learning techniques applied to the sentiment analysis of code-switched Tunisian dialect textual productions. We focus in particular on using LSTM-based RNNs.

3 Proposed Approach

Based on the study of various works of the literature addressing the SA of code switched text in different languages, and which have shown the effectiveness of approaches considering a multilingual classification of this kind of texts over those using several monolingual classifiers and requiring language detection, we propose to experiment this type of approach on TD. For this purpose, we propose to implement and evaluate some deep learning techniques that have not yet been explored in the context of TD SA. Our approach is based on RNNs (Recurrent Neural Networks) and more precisely on LSTM [17] networks.

Indeed, and as explained in [18], RNNs are considered to be more suitable for the SA task than CNNs (Convolutional Neural Networks), since they consider the sequential aspect of an input, where the words order is important and since they are able to treat inputs having variable lengths. Considering the remarkable results achieved in various classification tasks using LSTM networks, which are enhanced variants of RNNs [18], we propose to explore a set of models, based on variants of LSTM networks for the TD SA. The global architecture of the proposed models is presented in Fig. 1:

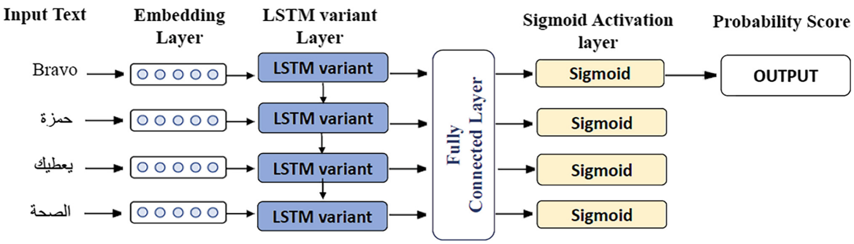

Fig. 1. Global architecture of the proposed models

In this architecture, each extracted review is mapped into vectors of real numbers, a very well-known technique when working with text called word embedding [19, 20]. In this technique, words are encoded as real-valued vectors in a high dimensional space, such as the similarity between the words in terms of close meaning, translates to closeness in the vector space. In our case, we used the **Keras**[2] Embedding Layer that is learned jointly within our examined LSTM models. Each word is taken as one input in a sequence. The first layer is initialized randomly, and the word embeddings are learned jointly on 13,655 reviews from the TSAC corpus.

In this work, we propose to explore the following LSTM-based models:

- **Long Short-Term Memory (LSTM).** LSTM is a type of artificial RNN, which was developed to address the issues of explosion and disappearance of gradients that can be experienced during the training of traditional RNNs [18]. A common LSTM unit consists of a cell, an input gate, an output gate and a forgetting gate. The cell stores values over arbitrary time intervals and the three gates control the flow of information into and out of the cell, controlling thus the information to forget or to pass on to the next time step.
- **Deep LSTM.** Also called stacked LSTM [21], it is an expansion of the original model that includes several hidden LSTM layers, each layer containing several memory cells. A stacked LSTM architecture can be indeed, defined as an LSTM model composed of several LSTM layers. An LSTM layer on the above provides a sequence output instead of a single value output to the LSTM layer below. The stacking of LSTM hidden layers allows the model to become deeper, by adding levels of abstraction of input observations over time and aims to give a more accurate description of the deep learning technique.
- **Bi-LSTM (Bidirectional LSTM).** A bidirectional LSTM consists of a forward LSTM, considering the input from the beginning to the end and backward LSTM, considering the input from the end to the beginning [18]. It interconnects two separate hidden layers that operate in opposite directions to a single output, which enables them to collect information from past and future states.
- **Deep (Stacked) Bi-LSTM.** This model is a combination of the two previous models. It is based on stacking two Bi-LSTM layers.

[2] https://keras.io/layers/embeddings/.

4 Experiments and Results

In this section we present the various performed experiments in order to evaluate the proposed models and discuss the obtained results. We first, start by describing the datasets we used in these experiments.

4.1 Used Datasets

For our experiments, we used the freely available annotated Tunisian dialect corpus, proposed by Mdhaffar et al. [5] and called TSAC (Tunisian Sentiment Analysis Corpus). TSAC is corpus of 17k comments that were extracted from Tunisian Facebook pages, mainly from official pages of Tunisian radios and TV channels during the period between January 2015 and June 2016. Extracted comments were cleaned and then manually labeled according to their polarity, into positive and negative comments. Some basic statistics describing the TSAC Corpus are given in Table 1.

Table 1. Statistics of the TSAC corpus [5]

	Positive	Negative
#Total words	63,874	49,322
#Unique words	24,508	17,621
#Comments	8,854	8,215
Avg. Sentence length	7.22	6.00

As shown, in Table 1, TSAC contains a total number of 17,069 TD comments in total of which, 8,854 are annotated as positive and 8,215 as negative comments. It contains both comments transcribed in the Arabic script and comments transcribed in the Latin script. Some comments mix the two scripts. As mentioned in [5], the collected corpus is composed of informal and non-standard vocabulary enclosing elongated words, abbreviations, emoticons, etc.

TSAC corpus was split into a training set and a test set as shown in Table 2.

Table 2. TSAC datasets

Data set	#Total comments
Total corpus	17,069
Training Set	13,655
Testing Set	1,707
Validation Set	1,707

We used the same datasets as in [5]; in order to be able to compare the performance of our proposed models with those experimented in [5].

4.2 Results and Discussion

All our experiments were performed using **Keras**[3], the Python Deep Learning library. The four proposed models were trained using the TSAC training set and tested using the TSAC test set (see Table 2).

Table 3 shows the obtained results while recalling the performance achieved by Mdhaffer et al. [5] models.

Table 3. Models' Evaluation Results on TSAC corpus

Mdhaffar et al. models [5]		Proposed models[a]	
Model	Accuracy	Model	Accuracy
SVM	0.77	LSTM	0.67
NB	0.58	Bi-LSTM	0.70
MLP	**0.78**	**Deep LSTM**	**0.90**
		Deep Bi-LSTM	0.88

[a]The models' parameters were tuned as follows: Epoch = 4 (all models); Batch size = 500 (LSTM model) and Batch size = 100 (the other models).

Let us first examine the evaluation results of the four proposed models which are based on various variants of LSTM networks. We can clearly see that, in our case the deep LSTM model leads to the best performance with 90% of correctly classified comments, while the simple LSTM model was the least performant with an accuracy of 67%.

We can also notice that using a bi-directional LSTM slightly improves the classification compared to the simple LSTM model. It seems that the Bi-LSTM model benefits only very little from exploring the preceding and following contexts in the case of the handled informal, multi-script and code-switched Tunisian dialect corpus. On the other hand, performance improved significantly with stacking layers within deep models. Deep LSTM and Deep Bi-LSTM had led indeed, to the best results with 0.9 and 0.88 accuracy values.

As the bi-LSTM model improves the simple LSTM classification and the deep-LSTM model significantly improves it, we were expecting the model combining layer stacking and bidirectional LSTM (the deep bi-LSTM model) to lead to the best results, but this was not the case since this model was outperformed by the deep LSTM model. Here again, the deep model doesn't seem to benefit from the double exploration of the preceding and following contexts (bi-LSTM). Even though predictions of complex deep neural networks are very difficult to interpret [22], we may explain this result by:

– The very particular nature of the handled data (composed of unformal vocabulary, mixing different scripts and enclosing code-switching that may lead to rather complex grammatical structures).

[3] https://keras.io/.

– The relatively limited size of the used corpus. We believe indeed, that the efficiency of the proposed models is likely to increase with a larger volume of training data.

When it comes to compare our results with those obtained by Mdhaffar et al. [5], the best results (90%) are obtained using the deep LSTM model, significantly outperforming the best results obtained by Mdhaffer et al. (78% and 77%, that were achieved by the MLP and the SVM classifiers).

5 Conclusion and Perspectives

In this paper, we tackled the issue of sentiment analysis of Tunisian dialect textual productions that are generated on social media. This type of data still presents number of challenges, due to its unformal, multi-scripted and code-switched nature.

We consider our work as a preliminary study that aimed to explore some deep learning techniques that were not yet applied to the case of the Tunisian dialect on the social media. We proposed indeed to experiment with various variants of RNNs, namely: LSTM, bi-LSTM, deep LSTM and deep Bi-LSTM. Our experiments performed on TSAC datasets [5] showed that we can reach a high performance with an accuracy of 90% using a deep (two-layer) LSTM model, outperforming the latest best proposed models on TSAC datasets in Mdhaffer et al. [5].

Furthermore, we intend to expand this work and enhance it, by working on generating a larger volume of annotated data on the one hand, and on the exploration of other approaches, such as introducing attention mechanism, using other types of RNNs such as GRUs, etc., on the other hand.

References

1. Poplack, S.: Code-switching (linguistic). Int. Encycl. Soc. Behav. Sci. **12**, 2062–4555 (2001)
2. Wang, Z., Zhang, Y., Lee, S., Li, S., Zhou, G.: A bilingual attention network for code-switched emotion prediction. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, pp. 1624–1634 (2016)
3. Younes, J., Souissi, E., Achour, H., Ferchichi, A.: Un état de l'art du traitement automatique du dialecte tunisien. TAL Traitement Automatique des Langues **59**, 93–117 (2018)
4. Medhaffar, S., Bougares, F., Estève, Y., Hadrich-Belguith, L.: Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In: Proceedings of the Third Arabic Natural Language Processing Workshop, Spain, pp. 55–61 (2017)
5. Mataoui, M.H., Zelmati, O., Boumechache, M.: A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. Res. Comput. Sci. **110**, 55–70 (2016)
6. Vilares, D., Alonso, MA., Gómez-Rodríguez, C.: Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Portugal, pp. 2–8 (2015)
7. Barman, U., Das, A., Wagner, J., Foster, J.: Code mixing: a challenge for language identification in the language of social media. In: Proceedings of the First Workshop on Computational Approaches to Code Switching, Qatar, pp. 13–23 (2014)

8. Supraja, C., Rao, V.M.: Emotion detection in code-switching text. Int. J. Eng. Technol. Sci. Res. **4**(12), 988–992 (2017)
9. Vilares, D., Alonso, MA., Gómez-Rodríguez, C. EN-ES-CS: an English-Spanish code-switching Twitter corpus for multilingual sentiment analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Slovenia, pp. 4149–4153 (2016)
10. Prabhu, A., Joshi, A., Shrivastava, M., Varma, V.: Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In: Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers, Osaka, pp. 2482–2491 (2016)
11. Baly. R., et al.: A characterization study of Arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In: Proceedings of the Third Arabic Natural Language Processing Workshop, Spain, pp. 110–118 (2017)
12. Nabil, M., Aly, M., Atiya, A.: ASTD: Arabic sentiment tweets dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Portugal, pp. 2515–2519 (2015)
13. Ameur, H., Jamoussi, S., Hamadou, A.B.: Exploiting emoticons to generate emotional dictionaries from Facebook pages. In: Czarnowski, I., Caballero, A., Howlett, R., Jain, L. (eds.) Intelligent Decision Technologies 2016, pp. 39–49. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39627-9_4
14. Sayadi, K., Liwicki, M., Ingold, R., Bui, M.: Tunisian dialect and modern standard Arabic dataset for sentiment analysis: Tunisian election context. In: Second International Conference on Arabic Computational Linguistics, ACLING, Turkey, pp. 35–53 (2016)
15. Mulki, H., Haddad, H., Bechikh Ali, C., Babaoglu, I.: Tunisian dialect sentiment analysis: a natural language processing-based approach. Computacion y Sistemas **22**(4), 1223–1232 (2018)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
17. Baziotis, C., Nikos, P., Christos D.: Datastories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, pp 747–754 (2017)
18. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of ICML, Helsinki, pp 160–167 (2008)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
20. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, pp. 6645–6649 (2013)
21. Lundberg, S., Lee, S.: Advances in Neural Information Processing Systems 30, pp. 4768–4777. Curran Associates, Inc. (2017)