Sentiment Analysis in Tunisian Social Media Using Deep Learning

Abir Messaoudi
iCompass, Tunisia
abir@icompass.digital

Hatem Haddad
iCompass, Tunisia

Chayma Fourati
iCompass, Tunisia
chayma@icompass.digital

1 Tunisian Sentiment Analysis

Tunisian dialect, is different from Modern Standard Arabic (MSA). In (Fourati et al., 2020), TUNIZI was referred to the Romanized alphabet used to transcribe informal Arabic for communication by the Tunisian Social Media community. (Younes et al., 2015) mentioned that 81% of the Tunisian comments on Facebook used Romanized alphabet. In (Abidi, 2019), a study was conducted on 1,2M social media Tunisian comments (16M words and 1M unique word) showed that 53% of the comments used Romanized alphabet while 34% used Arabic alphabet and 13% used code-switching script. The study mentioned also that 87% of the comments based Romanized alphabet are TUNIZI while the rest are French and English.

Previous work on Tunisian sentiment analysis (SA) were based on lexicon-based SA system (N., 2017) (Mulki et al., 2018), supervised SA system for Tunisian Arabic script tweets with different bag-of-word schemes used as features (Sayadi et al., 2016), document embeddings fed to train a Multi-Layer Perceptron (MLP) (Medhaffar et al., 2017), LSTM-based RNNs model (Jerbi et al., 2019) and syntax-ignorant n-gram embeddings representation composed and learned using an unordered composition function and a shallow neural model (Mulki et al., 2019).

More recently, Deep Neural Networks were widely used for this task, especially for the English language. In this paper, we explore the importance of various unsupervised word representations (word2vec, BERT) and we investigate the use of Convolutional Neural Networks and Bidirectional Long Short-Term Memory without using any kind of handcrafted features.

2 Proposed Approach

Recently, a sentiment analysis Tunisian Romanized alphabet dataset was introduced in (TUN) as ”TUNIZI”. For the purpose of this study, we filtred the TSAC dataset (Medhaffar et al., 2017) to keep only Tunisian Romanized comments. Datasets statistics are presented in Table 1.

We used three initial representations. The first being word2vec (Mikolov et al., 2013) which is a word-level embedding used in order to represent words in a high dimensional space. Then, frWaC (fau) used as a word2vec pretrained on 1.6 billion French word dataset. The use of the french pretrained word embedding is justified by the fact that most of the Tunizi comments present either french words or ones inspired by it. Finally, a Multilingual BERT (M-BERT) (Devlin et al., 2019). We decided to use the Bidirectional Encoder Representations from Transformers (BERT) as a contextual language model in its multilingual version as an embedding technique, that contain more than 10 languages including English and French. In this work, we experimented two classifiers, a Convolutional Neural Network (CNN) with different size of filters and a Bidirectional Long Short-Term Memory (Bi-LSTM), a variant of RNNs. These two models are followed by a fully connected layer and a softmax activation function for prediction. The different hyper-parameters achieving the best performances are shown in Table 2.

3 Preliminary Results

The word2vec representation trained on the TUNIZI dataset did not achieve better performances than frWaC representation trained on a French dataset. Indeed, frWac representation achieved a 0.702 accuracy performance when word2vec representation was only 0.672. This could be explained by the limited size of the TUNIZI dataset (82384

Dataset	#Words	#Uniq Words	#Comments	#Negative	#Positive	#Train	#Test
TUNIZI	82384	30635	9911	4679	5232	6908	2302
TSAC-TUNIZI	43189	17376	9196	3856	5340	7379	1817

Table 1: Datasets statistics

Embedding	Classifier	# filters
Word2vec	CNN	200
FrWaC	CNN	100
M-BERT	CNN	100
M-BERT	Bi-LSTM	-

Table 2: Hyper-parameters used during the training phase (number of epochs=3, batch size =16 for all experiments.)

Embedding	Classifier	Acc.	F1. macro
Word2vec	CNN	0.672	0.671
frWaC	CNN	0.702	0.690
M-BERT	Bi-LSTM	0.763	0.743
M-BERT	CNN	**0.783**	**0.781**

Table 3: Tunizi Classification Results.

words) used to train word2vec representation compared to the French dataset (1.6 billion words) used to train frWac representation. Training word2vec on the limited size of the TUNIZI dataset does not include all the Tunisian words, hence the out of vocabulary (OOV) phenomenon. The pretrained French word2vec did not solve the problem of OOV, but since it was trained on a very large corpus it handled most of French words that are frequent in TUNIZI comments.

Table 3 results suggest M-BERT embeddings combined with CNN or Bi-LSTM performed better then word2vec embeddings for all performance measures. This could be explained by the ability of Multilingual BERT to overcome the problem of switch-coding comments since it includes than 10 languages like English, French, and Spanish.

To confirm TUNIZI classification results, experiments were also performed on the TSAC-TUNIZI dataset with CNN and Bi-LSTM since they showed better performances than word2vec embeddings. From experiments presented in Table 4, we notice that the BERT embedding combined with the CNN leads to the best performance with a 93,2% of Accuracy compared to 92,6% scored by Bi-LSTM. This is also the case for the F1.macro performances with values of accuracy and F1.macro of 93,2% and 93%, respectively even though the TSAC-TUNIZI dataset is unbalanced having dominant positive polarity.

Classifier	Acc.	F1. macro
Bi-LSTM	0.926	0.925
CNN	**0.932**	**0.930**

Table 4: TSAC-TUNIZI Classification Results using M-BERT embedding

Experiments on TUNIZI and TSAC-TUNIZI datasets showed that M-BERT as an embedding technique outperforms word2vec representation. This could be explained by the switching code characteristic of the Tunisian dialect used on social media. Results suggest that M-BERT can overcome the out of vocabulary and switch code phenomena.

The CNN classifier performed better than the Bi-LSTM suggesting representation based Bi-LSTM did not benefit from the double exploration of the preceding and following contexts.

The obtained performances of our proposed approach were further compared against the baseline systems that tackled the TSAC dataset as shown in Table 5.

Model	Accuracy	F1.macro
(Medhaffar et al., 2017)	78%	78%
(Mulki et al., 2019)	86.5%	86.2%
(Jerbi et al., 2019)	90%	-
M-BERT+CNN	**93.2%**	**93%**

Table 5: TSAC dataset Compared Classification Results.

4 Future work

In this work, we have treated the Tunisian Romanized alphabet sentiment analysis task. We have experimented two different word-level representations (word2vec and frWaC) and two deep neural networks (CNN and Bi-LSTM).

A natural future step would involve releasing 3amBERT, a Tunisian version of BERT that should be learned on a very large and heterogeneous Tunisian dataset that can be applied to complex NLP tasks.

References

French word embeddings.

Tunizi dataset.

K. Abidi. 2019. Automatic building of multilingual resources from social networks : application to maghrebi dialects (doctoral dissertation). In *Université de Lorraine, France*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. In *AfricaNLP Workshop, Putting Africa on the NLP Map. ICLR 2020, Virtual Event*, volume arXiv:3091079.

M. Jerbi, H. Achour, and E. Souissi. 2019. Distributed representations of words and phrases and their compositionality. In *7th International Conference Arabic Language Processing: From Theory to Practice, Springer Cham., Nancy, France*, pages 122–131.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26, Curran Associates Inc., Lake Tahoe, Nevada*, page 3111–3119.

H. Mulki, H. Haddad, C. Bechikh Ali, and I. Babaoglu. 2018. Tunisian dialect sentiment analysis: A natural language processing-based approach. In *Computación y Sistemas*, pages 1223–1232.

H. Mulki, H. Haddad, M. Gridach, and I. Babaoglu. 2019. Syntax-ignorant n-gram embeddings for sentiment analysis of Arabic dialects. In *4th Arabic Natural Language Processing Workshop. Association for Computational Linguistics, Florence, Italy*, pages 30–39.

Karmani N. 2017. Tunisian arabic customer's reviews processing and analysis for an internet supervision system (doctoral dissertation). In *Sfax University, Sfax, Tunisia*.

K. Sayadi, M. Liwicki, R. Ingold, and M. Bui. 2016. Tunisian dialect and modern standard arabic dataset for sentiment analysis : Tunisian election context. In *2nd International Conference on Arabic Computational Linguistics, Konya, Turkey*.

J. Younes, H. Achour, and E. Souissi. 2015. Constructing linguistic resources for the tunisian dialect using textualuser-generated contents on the social web. In *15th International Conference Current Trends in Web Engineering, Rotterdam*.