Arabic dialect sentiment analysis with ZERO effort. Case study: Algerian dialect

Imane Guellil^{1,2,3}, Marcelo Mendoza⁴, Faical Azouaou¹

¹Laboratoire des Méthodes de Conception des Systèmes. Ecole nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. <http://www.esi.dz>

²School of Engineering and Applied Science (EAS), Aston University, Birmingham, UK. www.aston.ac.uk

³Folding Space Birmingham, UK.

⁴Department of Informatics, Universidad Técnica Federico Santa María, Santiago, Chile. <http://www.usm.cl>

i.guellil@aston.ac.uk, marcelo.mendoza@usm.cl, f_azouaou@esi.dz

Abstract This paper presents an analytic study showing that it is entirely possible to analyze the Arabic dialect's sentiment without constructing any resources. The idea of this work is to use the resources dedicated to a given dialect X for analyzing the sentiment of another dialect Y . The unique condition is to have X and Y in the same category of dialects. We apply this idea to the Algerian dialect, a Maghrebi Arabic dialect that suffers from limited available tools and other handling resources required for automatic sentiment analysis. To do this analysis, we rely on Maghrebi dialect resources and two manually annotated sentiment corpus for Tunisian and Moroccan dialect, respectively. We also use a large corpus for the Maghrebi dialect. We use a state-of-the-art system and propose a new deep learning architecture to classify the sentiment of the Algerian dialect automatically. Experimental results show that F1-score is up to 83%. It is achieved by Multilayer Perceptron (MLP) with Tunisian corpus and Long short-term memory (LSTM) with the combination of Tunisian and Moroccan. An improvement of 15% compared to its closest competitor was observed through this study. Ongoing work is aimed at manually constructing an annotated sentiment corpus for the Algerian dialect.

1 Introduction

Arabic is the official language of 27 countries with more than 400 million speakers. Arabic has two commonly used varieties, namely: 1) Modern Standard Arabic (MSA), which is used in formal written and verbal communication; and 2) Dialectal Arabic (DA), which is used in informal exchanges [23, 21]. Geographically, dialectal Arabic is classified into two main groups, namely Middle East (Mashriq) and North Africa (Maghrebi) dialects. Egyptian, Levantine, and Gulf dialects principally compose Mashriq dialects. Maghrebi dialects are principally composed by Algerian, Tunisian, and Moroccan dialects [45, 22]. Dialectal Arabic is abundantly present in social media and micro-blogging channels. Dialectal Arabic presents challenges for topical classifications and sentiment analysis [12, 16]. One example of such challenges is the lack of resources dedicated to these dialects. The majority of surveys done on Arabic dialect treatment shows that some dialects are more studied than others; for example, Egyptian dialect, Levantine, and Gulf [20]. Almost all the work done on Arabic dialects sentiment analysis begin by the resources construction (lexicon or corpus) dedicated to the studied dialect. Many studies on Arabic dialect treatment show that dialects that belong to the same group (Mashriq or Maghrebi) share enormous orthographic, lexical, and morphological characteristics.

In this context, this paper answers a crucial question: Instead of constructing new resources for handling each dialect separately; Why not take advantage of existing resources dedicated to similar dialects? To answer this

question, we focus on dialects that suffer from the limited available resources and studies such as the Maghrebi dialect. We use the resources presented in the literature for the Tunisian and Moroccan dialect to classify the sentiment of the Algerian dialect. Experimental results show that this technique improves the results presented in the literature by 15% for the F1-score. Our system achieves an F1-score of 83%, and it is up to 68% in the literature.

This paper is organized as follows. Section 2 presents the related work done in Arabic sentiment analysis. Section 3 introduces the datasets used in this study. The different systems used and developed in this work are discussed in Section 4. Section 5 presents the different results that we collected in this study. We present an analysis and a discussion about the presented results in Section 6. We conclude in Section 7, outlining future work.

2 Related work

Three main approaches are commonly used for sentiment analysis: the lexicon-based approach [41], machine learning (ML) based approach [33] and a hybrid approach [27]. ML-based sentiment analysis is the more dominant approach in the literature, but it requires annotated training data. To bridge the gap, almost all work in the literature focuses on annotating a set of data.

2.1 Annotated sentiment corpus

Among the corpora presented in the literature, we cite, OCA [39], AWATIF [1], LABR [6], TSAC [34], AraSenTi-Tweet [3], ElecMorocco2016 [14] and ASTD [36]. OCA contains 500 movie reviews collected from different Arabic web pages and blogs in Arabic (250 positives and 250 negatives). The reviews are manually preprocessed and segmented. AWATIF is a multi-genre corpus containing 10,723 Arabic sentences from three sources, namely: the Penn Arabic Treebank (ATB) [32], Wikipedia talk pages, and web forums. The sentences are manually annotated as objective or subjective, and subjective sentences are annotated as positive or negative [1]. LABR contains 63,257 Arabic comments annotated with stars ranging from 1 to 5. The authors considered positive comments with 4 or 5 stars, negatives with 1 or 2 stars, and neutrals with three stars. TSAC contains 17,060 comments (including 8215 positive and 8845 negative) in the Tunisian dialect annotated manually. AraSenTi-Tweet contains 17,573 Saudi tweets, manually annotated into four classes (positive, negative, neutral, and mixed). ElecMorocco2016 contains 10,254 manually annotated Facebook comments (including 3673 positive and 6581 negative). These comments are both in MSA and Moroccan dialect. ASTD contains 10,000 Arabic tweets annotated using Amazon Mechanical Turk as objective, subjective positive, subjective negative, or subjective mixed. However, almost all the corpora mentioned above were annotated manually, which is laborious. The purpose of this study is to rely on existing corpora created for dialects of the same family of the targeted dialect (the Maghrebi family, in our case).

2.2 Feature Extraction

The goal behind corpus annotation is sentiment classification. Sentiment classification is commonly carried out using various bag-of-words (BOW) models that are generated using feature selection methods. Bag of words (BOW) is commonly used to model text. BOW representations (e.g., Tf-Idf) are widely used in information retrieval mainly due to its simplicity as well as its efficiency in document recovery tasks. Despite its popularity, this approach has two significant drawbacks for sentiment analysis purposes: 1) Loss of word order in sentences (word interchangeability), and 2) Insufficient representation of semantic at word level [8].

Furthermore, to use a BOW representation in sentiment analysis, an appropriate word feature extraction process is required [2, 8]. Word and document embeddings are used as alternative representations of the previous ones [2, 13, 8, 5]. Al-Azani and El-Alfy [2] and Altowayan *et al.* [5] used a large-scale Arabic corpora to train a skip-gram word2vec model of text [35]. This corpus includes Quran-text (751,291 words), Arabic editions of international news networks such as CNN (24 million words) and BBC (20 million words), news articles based on a local Arabic newspaper (Watan-2004 with 106 million words) and finally a set of consumer reviews (40 million words). The authors used the generated word vectors to train different sentiment classifiers showing promising results in performance. Doc2vec sentence modelling [29] was used by Barhoumi [8] for sentiment classification on the corpus LABR [6]. The model proposed by these authors is composed of two parts: preprocessing part (in order to handle the input, light stemming) and classification part (to predict the polarity of the input). The authors used two classifiers, logistic regression (LR) and a multilayer perceptron (MLP). The input vector of the classifier is

the embedding obtained by learning a vector of the paragraph. This vector was a concatenation of the two learned vectors, one from a distributed memory version (DM) and the other from a distributed bag of words version (DBOW). Each model has 400 dimensions. This system was tested on the LABR corpus. However, the results obtained with this model are lower than those obtained in [6]. The authors affirm that the complexity of Arabic morphology requires more treatment than for another language such as English. El Mahdaouy *et al.* [13] also find results that allow affirming that the use of doc2vec improves the performance of classifiers in sentiment analysis. Recently, another algorithm for document representation known as FastText has emerged [26]. The performance of FastText is often compared with the performance obtained by word2vec in classification tasks [42, 40]. However, to the best of our knowledge, FastText has not been used yet in Arabic sentiment analysis. The purpose of this study is also to compare different feature extraction models in the case of Arabic dialects sentiment analysis.

2.3 Deep learning classification

Recently, deep learning algorithms such as Convolutional Neural Networks (CNN) [30], Long Short-Term Memory networks (LSTM) [24] or Bidirectional LSTM networks (Bi-LSTM) [15] have taken an important place in sentiment classification. In this context, Dahou *et al.* [9] introduce a method based on word2vec for Arabic sentiment classification, which evaluates polarity from product reviews. A convolutional neural network (CNN) model was trained on top of a set of pre-trained Arabic word embeddings. The authors used a multi-layer architecture defined by Kim *et al.* [28] to address this task. They applied their approach to different corpus presented in the literature studying LABR's performance, ASTD, ATT, HTL, and MOV, among others. More recently, Attia *et al.* [7] presented a language-independent model for multi-class sentiment classification using a simple multi-layer neural network architecture. This model contains five layers. The first layer is a randomly-initialized word embedding layer that turns sentences into a feature map. The second layer is a convolution neural network (CNN) layer that scans the feature map. The third layer is a global max-pooling applied to the output generated by the CNN layer to take the maximum score of each pattern. The purpose of the pooling layer is to reduce the dimensionality of the CNN representations by down-sampling the output and to keep the maximum value. The obtained scores are fed to a single feed-forward (fully-connected) layer with Relu activation. Finally, the output of that layer goes through a Softmax layer that predicts the output classes. As the proposed model is language-agnostic, it can be applied to multiple languages without using lexical resources as dictionaries or ontologies. The authors applied their model to sentiment classification in three languages: English, German, and Arabic. In the case of the Arabic language, the authors used the ASTD corpus [36] to validate the model.

3 Data Description

3.1 Tunisian dialect- TSAC corpus

TSAC (Tunisian Sentiment Analysis Corpus) contains 17060 user comments manually annotated to positive and negative polarities. This corpus contains comments written in Arabic scripts, Latin scripts known as Arabizi [10], and even a mixture of both. TSAC is a multi-domain corpus consisting of the text covering a maximum vocabulary from education, social and political domain [34]. In the context of this work, TSAC was filtered to keep only comments written in Arabic scripts. Hence, 7287 comments were selected, from which 3222 are positive, and 4065 are negative.

3.2 Morocco dialect- ElecMorocco2016

ElecMorocco2016 was extracted using Facebook Graph API¹ by targeting Moroccan on-line newspapers publishing news articles in the Arabic language. This corpus contains 10254 comments, from which 3673 are positive, and 6581 are negative. These comments are in both MSA and Morocco dialect and are focused on Moroccan's legislative elections, which took place on October 7, 2016 [14]. This corpus was pretreated where punctuation and special symbols (e.g., exclamation marks) were removed. In contrast with TSAC, this corpus contains only messages written in Arabic script.

¹<https://developers.facebook.com/docs/graph-api/>

3.3 Magrebi dialect- Northafrica corpus

North Africa corpus is a part of a voluminous Arabic Multi Dialect Text Corpora (containing 10,366,630 sentences) [4]. North Africa is the part of a corpus dedicated to the Maghrebi dialect. To construct this corpus, the authors target a set of the most used words for each dialect categories. For North Africa, 200 words (such as السيد meaning a school or يتجهم meaning to joke) were targeted. They also target a set of web pages dedicated to each dialect category. Afterward, the sentences containing the target words are extracted and pretreated.

4 Description of models

4.1 Classical machine learning algorithms

Three different representations, Bag of Words (BOW), word2vec [35] and doc2vec [29] are used. For BOW representation, CountVectorizer of the sklearn library is used. For word2vec, both continuous bag-of-words (CBOW) and skip-grams (SG) are used. For Doc2vec, the two methods presented in [29] are applied: 1) Distributed Memory Version of Paragraph Vector (PV-DM) and 2) Distributed Bag of Words Version of Paragraph Vector (PV-DBOW). For the classification part, different classifiers are used: 1) Naive Bayes (NB). 2) Logistic regression (LR). 3) Random Forest (RF). 4) Stochastic Gradient Descent (SGD) and 5) Linear Support Vector Classification (SVC). The algorithm developed in [5] is used for classification based on the word2vec model. Those developed in [17] are used for the doc2vec model. The algorithm developed in [5] relies on a large Arabic corpus. In this paper's context, the North Africa corpus is used because it is a large corpus written in the Maghrebi dialect.

4.2 Deep learning algorithms

Multi-Layer-Perceptrons (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (LSTM, and bi-LSTM) are used for sentiment classification. The first layer of each of these models is a word embedding layer that turns sentences into a feature map. FastText, with its two models CBOW and SG [25], is used to generate vectors of 300 dimensions. The architecture of each model is presented in Table 1. Adam optimizer is used for training. As a loss function, we use binary cross-entropy.

Classifier	Parameters
CNN	14 layers: 1 Embedding + 1 Convolutional+ 1 Max_pooling + 3 Convolutional + 1 Max_poolin + 3 Convolutional + 1 GlobalMaxPooling + 1 Dropout + 2 Dense layers
MLP	9 layers: 1 Embedding + 1 dense + 1 Activation + 1 dropout + 1 dense + 1 GlobalMaxPooling + 1 Dropout + 2 Dense layers
LSTM	9 layers: 1 Embedding + 2 LSTM + MaxPooling + 1 LSTM + 1 GlobalMaxPooling + 1 Dropout + 2 Dense layers
biLSTM	9 layers: 1 Embedding + 2 Bidirectional(LSTM) + MaxPooling + 1 Bidirectional(LSTM) + 1 GlobalMaxPooling + 1 Dropout + 2 Dense layers

Table 1: General architecture of the deep learning models used in this study

For all the used models, the embedding layer includes 6 parameters: *input_dim*, *output_dim*, *weights*, *input_length* and *trainable*. The *input_dim* represents the number of words that compounds the vocabulary (i.e., *input_dim* = 47,907 words for the corpus TSAC + ElecMorocco2016). An *output_dim* = 300 was used. The *weights* of the vectors constructed with fastText were used. For *input_length*, the sum of *mean* and *std* related to the training corpus was used (for example *input_length* = 22 for the corpus TSAC + ElecMorocco2016). Finally, *trainable* was set to False.

For the CNN model, 7 convolutional (*Conv1D*) layers were used. Each one of them has 64 *filters* and a *kernel_size* = 7. We also used a *relu* activation function and a *padding* = 'same'. The first Max_pooling layer represents a *MaxPooling1D* where the second one is a global max_pooling (i.e. *GlobalMaxPooling1D*). The rate of the *Dropout* was set to 0.5. The first *Dense* layer includes 32 units, it uses a *relu* activation with an *l2_regularizer* (with a *weight_decay* = 1e-4). The last *Dense* layer uses only 2 units (representing the two classes used, positive and negative) and a *sigmoid activation function*. In addition to the embedding layer with the same parameters used

in the CNN model, the MLP model has 4 dense layers with 64 units. The *relu* activation function is used with the first layers where the *sigmoid* one is used with last layer. These models also includes an activation layer with a *tanh* activation function. Both layers use dropout at 0.5. The LSTM model also relies on the same parameters for the embedding layer. It also includes 3 *LSTM* layers with 64 units and a *return_sequences* set to *True*. Same to the previous models, the rate of the dropout is 0.5. Finally, this model includes 2 *Dense* layers, where the first one has 32 units, a *relu activation function* and an *l2 regularizer*. The last *Dense* layer includes 2 units and uses a *sigmoid activation function*. The bi-LSTM model has the same configuration as the LSTM model. The main difference is that the bi-LSTM model use three-stacked bidirectional layers.

4.3 Set up

We carried out our experiments on three training corpora: (1) TSAC, (2) ElecMorocco2016, and (3) the combination between the two corpora TSAC + ElecMorocco2016. For the test corpora and to compare our work to others in the literature, we use the Algerian test corpus, which is manually annotated in [17]. This corpus contains 500 Facebook commentaries written in Arabic script (where 250 are positives, and 250 are negatives). For classification, we used two kinds of algorithms, shallow and deep learning algorithms. For shallow algorithms, we used the entire training corpus for the training phase. For deep learning classification, 80% of each corpus was reserved for the training phase, where the remaining 20% was used for validation.

5 Experimental results

5.1 Results on TSAC Dataset

Table 3 presents the performance of different shallow and deep classification algorithms using the TSAC dataset. Results show Precision (P), Recall (R), and F1-score (F1) for each vectorization method (BOW, word2vec, Doc2vec, and FastText). As Table 3 shows, the classifier associated with word2vec gives the best results in terms of precision, which is up to 89%. However, better recall and F1-score are achieved with a deep learning classifier, which is up to 83% for both metrics.

5.2 Results on ElecMorocco2016 Dataset

Table 4 presents the performance of the different classification algorithms by using the ElecMorocco2016 dataset. Table 4 shows that word2vec achieves the best results with a precision of up to 94%. The best recall is up to 86% with an NB classifier, and finally, the best F1-score is up to 78% with an SVC classifier. However, the results are very competitive for deep classifiers, where the F1-score with bi-LSTM is up to 77%.

5.3 Results on TSAC+ElecMorocco2016

Table 5 presents the performance of the different classification algorithms by combining the two datasets TSAC and ElecMorocco2016 for training. Table 5 shows that the classifier associated with word2vec gives the best results in terms of precision, which is up to 94%. However, the best recall and F1-score are achieved with a deep learning classifier (83% for both recall and F1-score).

6 Discussion and analysis

6.1 General analysis

Based on our results, our four main observations are: (1) The results on TSAC corpus are better than the results on ElecMorocco2016. (2) The results on the combination of both corpora do not introduce a significant improvement compared to the results obtaining on TSAC. (3) The best result in terms of F1 was achieved using deep learning models with an 83% F1-score. (4) For shallow classification, the best models are those based on word2vec.

TSAC is a multi-domain corpus, but ElecMorocco2016 is dedicated to the legislative area only. The test corpus used (annotated and used before in [17]) is also a multi-domain one, where the comments are extracted from different Algerian Facebook pages. For this reason, TSAC is more appropriate for the sentiment classification

Table 2: Detailed architecture of the deep learning models

Model	Layers	output shape	params
CNN	embedding_1	(None,22,300)	14372100
	conv1d_1	(None,22,64)	134464
	max_pooling1d_1	(None,11,64)	0
	conv1d_2	(None,11,64)	2873
	conv1d_3	(None,11,64)	2873
	conv1d_4	(None,11,64)	2873
	max_pooling1d_2	(None,5,64)	0
	conv1d_5	(None,5,64)	2873
	conv1d_6	(None,5,64)	2873
	conv1d_7	(None,5,64)	2873
	global_max_pooling1d_1	(None,64)	0
	dropout_1	(None,64)	0
	dense_1	(None,32)	2080
	dense_2	(None,2)	66
MLP	embedding_1	(None,22,300)	14372100
	dense_1	(None,22,64)	19264
	activation_1	(None,22,64)	0
	dropout_1	(None,11,64)	0
	dense_2	(None,11,64)	4160
	global_max_pooling1d_1	(None,64)	0
	dropout_2	(None,64)	0
	dense_3	(None,64)	4160
	dense_4	(None,2)	130
LSTM	embedding_1	(None,22,300)	14372100
	lstm_1	(None,22,64)	93440
	lstm_2	(None,22,64)	33024
	max_pooling1d_1	(None,11,64)	0
	lstm_3	(None,11,64)	33024
	global_max_pooling1d_1	(None,64)	0
	dropout_1	(None,64)	0
	dense_1	(None,32)	2080
	dense_2	(None,2)	66
BiLSTM	embedding_1	(None,22,300)	14372100
	bidirectional_1	(None,22,128)	186880
	bidirectional_2	(None,22,128)	98816
	max_pooling1d_1	(None,11,128)	0
	bidirectional_3	(None,11,128)	98816
	global_max_pooling1d_1	(None,128)	0
	dropout_1	(None,128)	0
	dense_1	(None,32)	4128
	dense_2	(None,2)	66

Type	Vec	Class	Metrics		
			P	R	F1
Shal-low	BOW	NB	0.74	0.73	0.73
		LR	0.81	0.80	0.79
		RF	0.78	0.73	0.72
		SGD	0.75	0.74	0.74
		SVC	0.79	0.79	0.79
	Word 2 vec	NB	0.71	0.62	0.66
		LR	0.86	0.71	0.78
		RF	0.78	0.63	0.70
		SGD	0.89	0.64	0.74
		SVC	0.85	0.71	0.77
	Doc 2 vec	NB	0.67	0.65	0.63
		LR	0.76	0.76	0.76
		RF	0.74	0.73	0.72
		SGD	0.78	0.77	0.77
		SVC	0.78	0.78	0.78
Deep	Fast text	CNN	0.81	0.80	0.81
		MLP	0.83	0.83	0.83
		LSTM	0.82	0.82	0.82
		biLSTM	0.81	0.81	0.81

Table 3: Classification results on TSAC dataset

task. It is important to note that even that ElecMorocco2016 is more voluminous than TSAC, TSAC gives the best results. The variety of training corpus is then more important than its size.

Word2vec vectorization is the best choice for shallow classification. This finding is so given that vectors were obtained from a sizeable Maghrebi corpus extracted in [4]. However, deep learning-based classification presents the more encouraging results where F1-score is up to 83%. It is also important to note that the use of FastText improves the results.

Our motivation when using the Algerian test corpus presented in [17] was to compare our results to the results in the literature. The authors' best results in [17] for test corpus written in Arabic script were up to 68% for F1-score. Hence, this work presents an improvement of 15% (where the best F1-score is up to 83%). These results were obtained using the resources offered for other dialects (Tunisian, Moroccan) similar to the studied dialect (Algerian dialect) in terms of orthography and morphology. No additional resources were constructed for the target dialect, saving time and human efforts.

However, these results can be improved by constructing an annotated corpus dedicated to the target dialect (Algerian dialect). The automatic construction of this corpus was investigated in [17], but the results obtained by these authors were not very promising. Hence, the manual construction of an Algerian dialect sentiment corpus is necessary. For this process, following the guideline proposed in the construction of TSAC represents a good starting point.

6.2 The improvements to integrate to this study

In order to extract features from the used corpora, we opt for two well-known and largely used word embedding models being Word2vec and fastText. For document embedding, we opt for Doc2vec. However, more recently, a large panoply of models was proposed. In this context, two families of models have been proposed: 1) Contextual word embeddings and 2) Transformer word embeddings. The first model introduced in the family of contextual models is *Elmo* [38]. Unlike traditional word embeddings such as word2vec and fastText, the Elmo assigned to a token or word is a function of the entire sentence containing that word. Therefore, the same word can have different word vectors in different contexts. This approach allows improving the state-of-the-art results related to different NLP tasks, including sentiment analysis. Concerning the transformers, different models were proposed including BERT [11], [31] and XLNet [44]. The idea of transformers was first introduced by Vaswani et al. [43]. The

Type	Vec	class	Metrics		
			P	R	F1
Shal-low	BOW	NB	0.65	0.64	0.63
		LR	0.77	0.68	0.65
		RF	0.72	0.57	0.49
		SGD	0.72	0.66	0.63
		SVC	0.73	0.71	0.70
	Word 2 vec	NB	0.60	0.86	0.71
		LR	0.93	0.61	0.74
		RF	0.71	0.46	0.56
		SGD	0.94	0.61	0.74
		SVC	0.90	0.68	0.78
	Doc 2 vec	NB	0.69	0.68	0.67
		LR	0.78	0.75	0.74
		RF	0.72	0.66	0.64
		SGD	0.76	0.76	0.76
		SVC	0.76	0.74	0.73
Deep	Fast text	CNN	0.77	0.76	0.76
		MLP	0.81	0.76	0.75
		LSTM	0.81	0.75	0.74
		biLSTM	0.78	0.77	0.77

Table 4: Classification results on ElecMorocco2016

Type	Vec	class	Metrics		
			P	R	F1
Shal-low	BOW	NB	0.69	0.68	0.68
		LR	0.83	0.81	0.80
		RF	0.82	0.77	0.76
		SGD	0.79	0.78	0.78
		SVC	0.79	0.79	0.79
	Word 2 vec	NB	0.58	0.73	0.65
		LR	0.91	0.74	0.81
		RF	0.81	0.62	0.70
		SGD	0.94	0.68	0.79
		SVC	0.90	0.72	0.80
	Doc 2 vec	NB	0.75	0.69	0.67
		LR	0.82	0.81	0.81
		RF	0.77	0.75	0.74
		SGD	0.81	0.78	0.77
		SVC	0.80	0.79	0.79
Deep	Fast text	CNN	0.82	0.81	0.81
		MLP	0.83	0.81	0.81
		LSTM	0.84	0.83	0.83
		biLSTM	0.82	0.80	0.80

Table 5: Classification results on Tsac+ElecMorocco2016 dataset

transformer architecture reads entire sequences of tokens at once. In a sense, the model is non-directional, while LSTMs read sequentially (left-to-right or right-to-left). The attention mechanism allows for learning contextual relations between words. Hence, our future works will be dedicated to integrating these emerging models to improve the results.

This paper's idea was principally inspired by transfer learning [37], where we used the model trained on a given dialect on a dataset of another dialect. As we have only a reduced Algerian annotated dataset for our experiments, we were unable to fine-tune our models. Hence, to improve the results of the proposed models, we are also working on extending the number of samples used for the Algerian dialect.

7 Conclusion and Perspectives

In this paper, we present an analytical study dedicated to automatically analyzing the Arabic dialect's sentiment by relying on the resources constructed for another similar dialect. We apply this idea to the Algerian dialect, which shares many characteristics with other Maghrebi dialects such as: Tunisian and Moroccan. We also use different classification systems (shallow and deep). The better results were achieved with deep learning in terms of F1-score.

This study represents a start point for our future work, where we plan to construct a manually annotated sentiment corpus for the Algerian dialect. We will compare that results with those obtained in this study. We also plan to integrate Arabizi messages ². For handling Arabizi, we will firstly rely on the transliteration process [18, 19], in order to transform Arabizi into Arabic. Additionally, enlarging the corpus with Algerian dialect (with both Arabic and Arabizi script) will allow us to analyze the sentiment of the Maghrebi dialect. Finally, we also plan to construct different sentiment lexicons for each of these dialects and integrate the lexicon-based approach in this study.

Acknowledgments

Mr. Mendoza acknowledge funding from the Millennium Institute for Foundational Research on Data. Mr. Mendoza was also funded by ANID PIA/APOYO AFB180002 and ANID FONDECYT grant 1200211.

References

- [1] Muhammad Abdul-Mageed and Mona T Diab. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer, 2012.
- [2] Sadam Al-Azani and El-Sayed M El-Alfy. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, 109:359–366, 2017.
- [3] Nora Al-Twairish, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72, 2017.
- [4] Khalid Almeman and Mark Lee. Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In *Communications, signal processing, and their applications (iccspa), 2013 1st international conference on*, pages 1–6. Citeseer, 2013.
- [5] A Aziz Altowayan and Lixin Tao. Word embeddings for arabic sentiment analysis. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3820–3825. IEEE, 2016.
- [6] Mohamed Aly and Amir Atiya. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 494–498, 2013.
- [7] Mohammed Attia, Younes Samih, Ali El-Kahky, and Laura Kallmeyer. Multilingual multi-class sentiment classification using convolutional neural networks. In *LREC*, 2018.

²Arabic words written with Latin letters

- [8] Amira Barhoumi, Yannick Estève, Chafik Aloulou, and Lamia Hadrich Belguith. Document embeddings for arabic sentiment analysis. 2017.
- [9] Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2418–2427, 2016.
- [10] Kareem Darwish. Arabizi detection and conversion to arabic. *arXiv preprint arXiv:1306.6755*, 2013.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [12] Rehab M Duwairi. Sentiment analysis for dialectal arabic. In *Information and Communication Systems (ICICS), 2015 6th International Conference on*, pages 166–170. IEEE, 2015.
- [13] Abdelkader El Mahdaoui, Eric Gaussier, and Saïd Ouatik El Alaoui. Arabic text classification based on word and document embeddings. In *International Conference on Advanced Intelligent Systems and Informatics*, pages 32–41. Springer, 2016.
- [14] Abdeljalil Elouardighi, Mohcine Maghfour, Hafdalla Hammia, and Fatima-zahra Aazi. A machine learning approach for sentiment analysis in the standard or dialectal arabic facebook comments. In *Cloud Computing Technologies and Applications (CloudTech), 2017 3rd International Conference of*, pages 1–8. IEEE, 2017.
- [15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.
- [16] Imane Guellil, Ahsan Adeel, Faical Azouaou, Fodil Benali, Ala-eddine Hachani, and Amir Hussain. Arabizi sentiment analysis based on transliteration and automatic corpus annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 335–341, 2018.
- [17] Imane Guellil, Ahsan Adeel, Faical Azouaou, and Amir Hussain. Sentialg: Automated corpus annotation for algerian sentiment analysis. In *9th International Conference on Brain Inspired Cognitive Systems (BICS 2018)*, 2018.
- [18] Imane Guellil, Faical Azouaou, Mourad Abbas, and Sadat Fatiha. Arabizi transliteration of algerian arabic dialect into modern standard arabic. In *Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation*, 2017.
- [19] Imane Guellil, Faical Azouaou, Fodil Benali, ala-eddine Hachani, and Houda Saadane. Approche hybride pour la translittération de l’arabizi algérien : une étude préliminaire. In *Conference: 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, France*. *FranceAt: Rennes, France*. https://www.researchgate.net/publication/326354578_Approche_Hybride_pour_la_translitteration_de_l_arabizi_algerien_une_etude_preliminaire, 2018.
- [20] Imane Guellil and Azouaou Faical. Bilingual lexicon for algerian arabic dialect treatment in social media. In *WiNLP: Women & Underrepresented Minorities in Natural Language Processing (co-located with ACL 2017)*. http://www.wlnlp.org/wp-content/uploads/2017/final_papers_2017/92_Paper.pdf, 2017.
- [21] Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 2019.
- [22] Imène Guellil and Faical Azouaou. Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect. In *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, pages 724–731. IEEE, 2016.

- [23] Salima Harrat, Karima Meftouh, and Kamel Smaïli. Machine translation for arabic dialects (survey). *Information Processing & Management*, 2017.
- [24] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [25] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. Fast-text.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.
- [26] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
- [27] Aamera ZH Khan, Mohammad Atique, and VM Thakare. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCE)*, page 89, 2015.
- [28] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
- [29] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.
- [30] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1995.
- [31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [32] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo, 2004.
- [33] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.
- [34] Salima Medhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, 2017.
- [35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [36] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, 2015.
- [37] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [38] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
- [39] Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. Oca: Opinion corpus for arabic. *Journal of the Association for Information Science and Technology*, 62(10):2045–2054, 2011.
- [40] Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *arXiv preprint arXiv:1808.09238*, 2018.

- [41] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
- [42] Shabnam Tafreshi and Mona Diab. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913, 2018.
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.
- [45] Omar F Zaidan and Chris Callison-Burch. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202, 2014.