Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 246–253

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects

Mohamed Lichouri[a,b,*], Mourad Abbas[a], Abed Alhakim Freihat[c], Dhiya El Hak Megtouf[a]

[a] *Computational Linguistics Department, CRSTDLA, Algeria*
[b] *USTHB University, Algeria*
[c] *University of Trento, Italy*

Abstract

In this paper, we investigate a set of methods for textual Arabic Dialect Identification, where we considered word-level and sentence-level approaches. We used three classifiers, namely: Linear Support Vector Machine L-SVM, Bernoulli Naive Bayes BNB and Multinomial Naive Bayes MNB. Then we combined them by using a voting procedure. We carried out experiments on two sets of dialects: the first one, PADIC, which consists of parallel sentences in Maghrebi and Middle Eastern dialects; and the second, a set of Algerian dialects only, that we built manually. For the Arabic dialects, we obtained an average accuracy of 92%. For Algerian dialects, our approach yielded an average accuracy of about 76%.

Keywords: Sentence-Level Dialect Identification; Word-Level Dialect Identification; Naive Bayes Classifier; SVM; Algerian Dialects; Arabic Dialects

1. Introduction

Dialect Identification (DI) is one of the important topics in Natural Language Processing (NLP) with an ever-growing number of work. Dialectal Arabic is an open problem which attracts many researchers. That is why recently, a couple of work have been done, as Spoken Arabic Dialect Identification [1, 2, 3, 4, 5], Collecting Arabic Dialect Corpora [6, 7, 8], Arabic Dialect Machine Translation [9, 10], Text Diacritization in Arabic Dialect [11]. There is also a couple of work done on the textual data, like Arabic Dialect Word Segmentation [12, 13] and Arabic Textual Dialect Identification [6, 14, 15, 16, 17]. Through all these works, it has been shown that Dialectal Arabic (DA) is a problem in itself because it differs strongly from Modern Standard Arabic (MSA) as well as the number of variations in the same and different Arabic Countries. But on another view, Arabic Dialect becomes an important source of information

* Corresponding author
E-mail address: licvol@gmail.com, mlichouri@usthb.dz

1877-0509
10.1016/j.procs.2018.10.484

for a couple of fields: social, security and market studies. As an example, if a Tech-Company wants to know for which Arabic country a certain product is successful or not, they must analyze its social web page comments (where by default most reviewers will not determine where they are from). So a need for a Arabic Dialect identification process is in need here. This paper is organized as follow: related work is presented in section 2, whereas in section 3 we present our approach for Dialect Identification based on word-level and sentence-level. In section 4, we present the different corpora used in our experiments. In section 5, we summarize the results and discussions. Finally, we present our conclusions and perspectives in section 6.

2. Related Work

Currently most of works related to Arabic Text Dialect Identification are based mainly on Egyptian, Iraqi and Levantine dialects in addition to MSA. Elfardy et al. [14] introduced a supervised approach for performing sentence level dialect identification between MSA and Egyptian dialects with an accuracy of over 85% in an Arabic online-commentary dataset. An accuracy of 89.1% on this same dataset by using a linear support-vector machine was achieved by Tillmann et al. [18]. Whereas Sadat et al. [15] described the usage of probabilistic models across social media datasets using the character n-gram Markov language model and Naive Bayes classifiers which achieved an overall accuracy of 98%. Another effort was done by Durandin et al. [16] to reduce time expenses on manual annotation of data from social media by proposing the use of two corpora, a small manually annotated corpus and a big one which was grabbed from the Web automatically using word-marks. These marks have permitted the authors to achieve an accuracy of 92% for four Arabic dialects (Levantine, Egyptian, Saudi and Iraqi). There is also the work done by Malmasi et al. [19] in the Discriminating Similar Languages (DSL) 2015 shared task, where they constructed a classifier ensemble composed of several Support Vector Machines (SVM) based classifiers, each trained on a single feature type achieving the best accuracy of 95.54%. We note also the work presented in [17], where a basic DI on PADIC corpus (which consists of parallel sentences in Maghrebi and Middle Eastern dialects) has been conducted, and yielded an overall accuracy of 56%. Al-Badrashiny et al. [20] suggested a new approach which considers a hybrid approach for performing token and sentence levels Dialect Identification in Arabic. They tried to identify whether each token in a given sentence belongs either to Modern Standard Arabic or Egyptian Dialectal Arabic (EDA) and whether the whole sentence is mostly EDA or MSA. They achieved an overall accuracy of over 90%. In our paper, we further considered this information at another level, trying to answer this question: If a sentence has more tokens belonging to dialect x than dialect y, how can we decide that this sentence belongs to the dialect x or y, and how we explain this decision. The answer to this question is shown in Subsection 3.1.

3. Dialects Identification Model

Our approach is inspired by the work of Al-Badrashiny et al. [20]. Indeed, we address the problem of dialect identification for Arabic and Algerian on the word and sentence levels respectively. The source of our used dataset is the PADIC corpus [7, 9], in addition to a set of dialectal Algerian texts written by people from eight Algerian cities. We first classify each word in the input sentence by assigning one of the tags defined in table 1 for Arabic and Algerian Dialects. The fully tagged words in the given sentence are then used to determine the class of this sentence, according to the tags defined in table 1 and which are fully described in subsections 4.1 and 4.2.

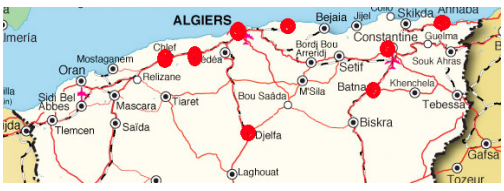

Fig. 1: Location of Algerian Dialects

Dialects	Tags
Arabic	MSA, ALG, PAL, TUN, SYR, MOR
Algerian	ALG, CST, TNS, JLF, KAB, ANB, BTN, DFL

Table 1: Arabic and Algerian dialects

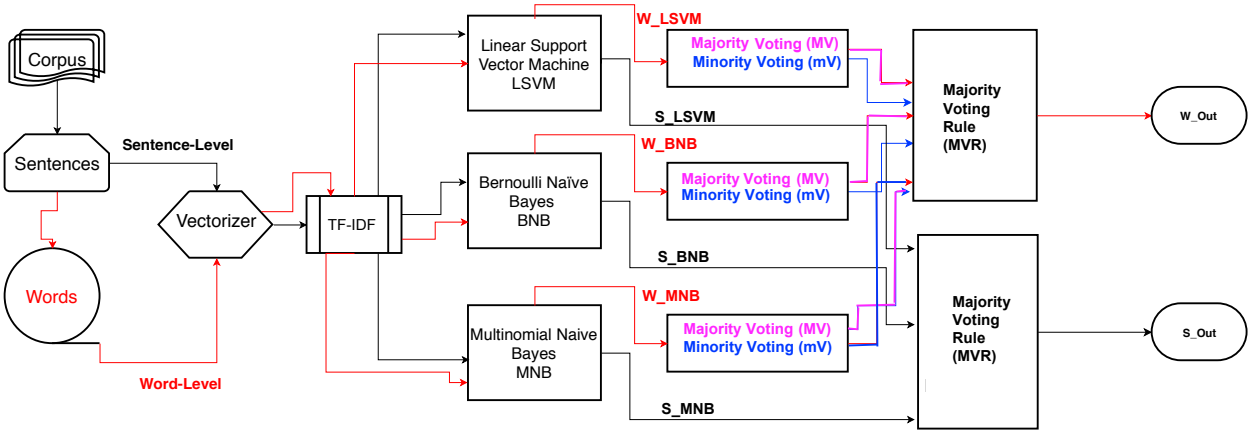

Fig. 2: Word (Red) vs Sentence (Black) Level Dialect Identification

3.1. Word Level Identification (WLI)

In order to identify the class of a token in a given sentence, we used the classifiers (Linear Support Vector Machine L-SVM, Bernoulli Naive Bayes BNB, Multinomial Naive Bayes MNB). These classifiers are trained on the dataset mentioned in Table 1. Figure 2 shows the pipeline of the word level identification process in red color. The pipeline consists of three main pathways with some pre-processing components. All the sentences will be converted to a bag-of-words matrix then transformed to a weighted matrix by the Tf-Idf technique. Each word W_i in a sentence S_j is assigned to its category C_{ij}^k, by all of the classifiers, where k stands for the k^{th} category (dialect). A Majority and a Minority voting methods are applied to the three outputs ($W\_LSVM = \{C_{ij}^k\}$), ($W\_BNB = \{C_{ij}^k\}$), ($W\_MNB = \{C_{ij}^k\}$) (see Figure 2) to extract the category of each sentence, based on the words' categories. In fact, we justify the use of such methods by the following two questions:

- Does applying the majority voting (MV) at the word level lead to the correct category. In other words, is it correct to assign a category C (dialect) to a sentence S if more than 50 % of its words belong to this category?
- If we take into consideration the catchwords of a dialect where they appear only once in a sentence, like "Bezaf" (Algeria), "Barsha" (Tunisia), "Ouakha" (Moroccan). Is it the best decision to consider the minority voting (mV) approach, i.e. to label this sentence with the dialect tag to which belongs the catchword?

Let us consider the example of the sentence "**shryt ktab altrjmh ghaly bzaf**". If we apply the MV rule, then the sentence is recognized as an MSA sentence since the words "**ktab**","**altrjmh**" and "**ghaly**" belong to MSA. However, the sentence is dialectal because it contains the word "**bzaf**" which can be found only in Algerian dialect.

The final step of the pipeline, is to make a combination of the three classifiers using Majority Voting Rule (MVR),giving the final decision for the word-level identification approach (see algorithm 1).

3.2. Sentence Level Identification (SLI)

We have considered the same steps followed in the Word-level identification approach. Figure 2 shows the pipeline of the sentence level identification component in black. After achieving pre-processing, we carried out identification using the three aforementioned classifiers. Finally, the classifiers outputs are combined by the Majority Voting (MV) method (see algorithm 2).

Algorithm 1 Word Level Identification

procedure WLI(*corpus*)
  Preparing Train and Test Data ▷ (Step 1)
  Pre-Processing Phase ▷ (Step 2)
  Training Phase for LSVM, BNB and MNB ▷ (Step 3)
  for $S \in Test$ **do** ▷ Testing Phase (Step 4)
    Initiate W_LSVM, W_BNB , W_MNB
    for $W \in S$ **do**
      Predicting the Class by LSVM, BNB and MNB
      Updating W_LSVM, W_BNB and W_MNB
  Applying (MV) to the 3 outputs: W_LSVM, W_BNB and W_MNB
  Applying (mV) to the 3 outputs: W_LSVM, W_BNB and W_MNB
  Final Result ← Combination of the six outputs using Majority Voting. ▷ Final Phase (Step 5)

Algorithm 2 Sentence Level Identification

procedure SLI(*corpus*) ▷ Same Step 1 to 3 in WLI
  Execute Step 1 to 3
  for $S \in Test$ **do** ▷ Testing Phase (Step 4)
    Predicting the Class by LSVM, BNB and MNB
    Updating the 3 outputs: S_LSVM, S_BNB and S_MNB
  Final Result ← Combination of the three outputs using Majority Voting. ▷ Final Phase (Step 5)

4. Dataset

4.1. Arabic Dialects Dataset

For the Arabic Dialect Identification we used the PADIC corpus [7, 9] in the training and test phase. We should note that PADIC [1] is a multi-dialectal corpus built in the framework of the National Research Project "TORJMAN" [2], led by the Research Center for Arabic Language and funded by the Algerian Ministry of Higher Education and Scientific Research. The whole corpus is described in Table 2.

Arabic Dialects		ALG (Algeria)	TUN (Tunisia)	MOR (Morocco)	SYR (Syria)	PAL (Palestine)	MSA
Train (PADIC)	# Sentences	5,130	5,130	5,130	5,130	5,130	5,130
	# Words	28,111	31,189	33,987	31,961	33,749	35,413
Test (PADIC)	# Sentences	1,284	1,284	1,284	1,284	1,284	1,284
	# Words	7,122	7,687	8,549	8,158	8,405	8,732
External dataset	# Sentences	2,323	2,014	2,970	1,812	1,053	492
	# Words	27,610	25,284	29,892	10,543	6,589	3,850

Table 2: The Arabic Dialects corpus (PADIC).

[1] https://sourceforge.net/projects/padic/
[2] https://sites.google.com/site/torjmanepnr/

4.2. Algerian Dialects Dataset

To the best of our knowledge, there is no Gold Algerian corpora dedicated for Dialect Identification. Our objective is to build such a resource and to cover the large spectrum of Algerian dialects which are very different from one region to another. Thus, we started to build textual data from eight Algerian cities ranging from the east to the west. We should note that dialects used by people living in the eastern cities are close to Tunisian, and those used by people in the west of the country are close to Moroccan. Besides, there are other Algerian dialects which are very different, like Berber dialects (Spoken in Tizi-Ouzou and other cities). We selected a 100 sentences (in MSA) from PADIC corpus, and translated them to the eight following dialects (Tenes "TNS", Constantine "CST", Djelfa "JLF", Ain-Defla "DFL", Tizi-Ouzou "KAB", Batna "BTN", Annaba "ANB" and Algiers' Dialect "ALG" -see figure 1-). The specification of this corpus is presented in Table 3.

Algerian Dialects		ALG (Algiers)	CST (Constantine)	TNS (Tenes)	JLF (Djelfa)	KAB (Tizi Ouzou)	ANB (Annaba)	BTN (Batna)	DFL (Ain Defla)
Train (Collected)	# Sentences	80	80	80	80	125	80	36	80
	# Words	659	314	632	601	683	670	289	653
Test (Collected)	# Sentences	20	11	20	20	32	20	9	20
	# Words	156	86	149	147	173	160	62	166

Table 3: The Algerian Dialects corpus consisting of eight dialects

5. Results and Discussion

As mentioned above, we used PADIC corpus for training and testing. At the sentence level, the classifiers (LSVM, BNB and MNB) are trained, using the aforementioned corpora, then a testing process has been conducted. At the word level, we introduced two methods to boost the three classifiers performance, namely Majority and Minority Voting approaches. We display in Table 4, results for both word-level and sentence-level based identification for the different pairs of dialects. It should be noted that at word level, the accuracy of the LSVM, BNB and MNB classifiers are obtained by applying both the Majority and the Minority voting.
Furthermore, it is worth mentioning that the combination of the classifiers outputs, by the Majority voting rule, at both word and sentence levels has been applied, and mentioned as (WL_MV) and (SL_MV) respectively.

5.1. Dialectal Arabic Identification Results

As shown in table 4, for binary classification, we noticed that: a) the best classifier for all the dialects is the Bernoulli Naive Bayes (BNB); b) the best results are obtained for the language pairs MSA-Maghrebi compared to MSA-Levantine; This can be explained by the fact that Syrian and Palestinian are closer to MSA than Maghrebi dialects are.
It is worth mentioning that the word-level approach performance is lower than the sentence-level one, nevertheless it outperforms results shown in [17] where the same corpus had been used "PADIC". The highest accuracy of 96.21% is obtained for the ALG-MSA language pair by the MNB classifier. In addition, we performed a multiclass classification (the 5 dialects and MSA), the best accuracy is 73.15%.

5.2. Dialectal Algerian Identification Results

We summarize in Table 5 all the results obtained with the Algerian corpus. The performances of the used classifiers at both sentence and word levels are not very different from each other, compared to the ones recorded for Arabic dialects. In fact the average accuracy for Algerian dialects is ranging between 62% to 76% - see Figure 3. The best results are obtained for the following pairs: ALG-BTN (96.15%), ANB-BTN (100%), KAB-TNS (98.07%), TNS-BTN (100%), CST-BTN (100%), BTN-DFL (96.55%) and JLF-KAB (94.23%). We should note that the word level approach yielded better results for Algerian dialects than for Arabic dialects. It is worth mentioning that the Minority

Language Pairs	Word-Level DI (%)							Sentence-Level DI (%)			
	LSVM (MV)	LSVM (mV)	BNB (MV)	BNB (mV)	MNB (MV)	MNB (mV)	WL_MV	LSVM	BNB	MNB	SL_MV
ALG-MSA	87.75	47.52	87.32	48.49	87.60	48.22	48.30	94.69	96.02	**96.21**	**96.21**
ALG-TUN	76.76	44.09	72.82	49.20	72.86	49.23	50.95	85.06	87.99	87.01	87.29
ALG-MOR	79.84	42.80	79.68	42.88	79.84	42.80	42.80	88.61	**91.22**	90.79	90.83
ALG-PAL	84.32	48.14	84.09	48.65	84.09	48.57	48.69	91.92	**93.80**	93.64	93.72
ALG-SYR	84.24	45.92	85.57	45.57	85.61	45.53	56.25	92.63	**94.69**	94.50	94.58
TUN-MSA	81.32	52.01	81.20	52.55	81.36	52.04	52.47	91.46	**94.38**	93.56	93.60
TUN-MOR	78.51	45.92	77.77	47.09	78.20	46.51	46.90	87.75	**90.25**	89.35	89.47
TUN-PAL	78.16	50.79	76.92	52.47	77.30	51.85	51.81	91.38	**93.68**	92.94	93.13
TUN-SYR	80.77	46.74	80.07	47.40	80.11	47.40	47.52	90.79	**92.55**	92.24	92.51
MOR-MSA	85.84	43.82	84.13	42.06	84.13	42.06	49.47	92.63	**94.65**	93.91	93.99
MOR-PAL	84.01	42.41	84.05	42.45	84.01	42.41	41.13	91.11	**93.45**	92.98	93.33
MOR-SYR	82.26	46.51	81.59	47.71	81.94	47.09	50.33	91.73	**93.80**	93.21	93.48
PAL-MSA	76.17	43.89	75.12	45.57	75.82	44.24	45.10	86.39	**89.86**	88.77	89.04
PAL-SYR	66.15	46.82	66.08	46.78	66.08	46.78	50.33	79.61	**81.59**	80.11	80.58
SYR-MSA	78.63	46.08	78.01	47.29	78.20	46.97	47.09	90.64	**92.16**	92.12	**92.16**
Multiclass	52.02	21.20	50.21	22.02	51.15	21.48	32.75	70.80	**73.15**	72.63	72.61

Table 4: Word vs Sentence Level Dialect Identification Applied to Arabic Dialect Corpus. MV and mV stand for Majority and minority Voting. WL_MV and SL_MV stand for Combination of the three classifiers at word and sentence levels respectively.

voting rule performed well for Algerian while Majority voting provided better results for Arabic dialects. The reason might be that the studied Algerian dialects are characterized by their own catchwords, which are distinct from each other. .

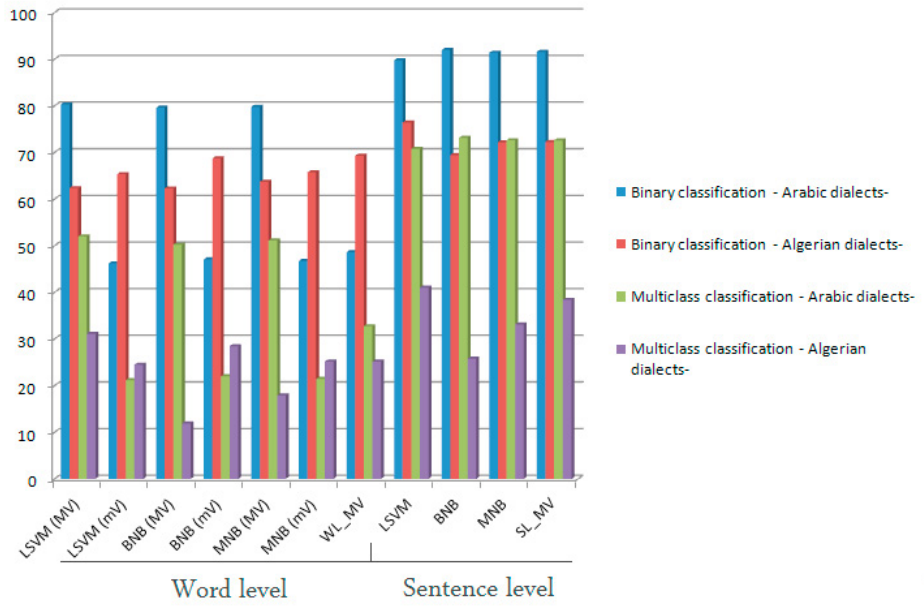

Fig. 3: Average accuracy for Arabic and Algerian dialects at word and sentence level -Binary vs Multiclass classification.

Language Pairs	Word-Level DI (%)							Sentence-Level DI (%)			
	LSVM (MV)	LSVM (mV)	BNB (MV)	BNB (mV)	MNB (MV)	MNB (mV)	WL_MV	LSVM	BNB	MNB	SL_MV
ALG-ANB	30.00	62.50	30.00	62.50	30.00	62.50	62.50	57.50	47.50	42.50	50.00
ALG-KAB	71.15	59.61	71.15	59.61	71.15	59.61	59.61	**94.23**	84.61	**94.23**	**94.23**
ALG-JLF	55.00	42.50	55.00	42.50	55.00	42.50	42.50	65.00	75.00	65.00	65.00
ALG-TNS	32.50	60.00	32.50	57.50	32.50	60.00	57.50	52.50	35.00	40.00	40.00
ALG-CST	77.41	58.06	80.64	54.83	77.41	54.83	58.06	74.19	77.41	74.19	74.19
ALG-BTN	75.86	89.65	75.86	**96.55**	75.86	89.65	**96.55**	89.65	68.96	82.75	82.75
ALG-DFL	12.50	70.00	12.50	70.00	12.50	70.00	70.00	15.00	12.50	12.50	12.50
ANB-KAB	76.92	61.53	76.92	61.53	76.92	61.53	61.53	**94.23**	75.00	92.30	92.3
ANB-JLF	35.00	62.50	35.00	65.00	35.00	62.50	65.00	**90.00**	85.00	85.00	85.00
ANB-TNS	45.00	62.50	45.00	62.50	45.00	62.50	62.50	55.00	50.00	52.50	52.50
ANB-CST	67.74	48.38	67.74	64.51	67.74	74.19	64.51	74.19	64.51	64.51	64.51
ANB-BTN	51.72	96.55	51.72	75.86	51.72	**100.00**	75.86	82.75	55.17	72.41	72.41
ANB-DFL	40.00	65.00	42.50	60.00	40.00	65.00	62.50	45.00	45.00	40.00	42.50
KAB-JLF	51.92	71.15	51.92	71.15	51.92	71.15	71.15	**94.23**	78.84	**94.23**	**94.23**
KAB-TNS	67.30	80.76	69.23	65.38	69.23	65.38	84.61	96.15	88.46	**98.07**	96.15
KAB-CST	76.19	92.85	71.42	**95.23**	76.19	92.85	90.47	**95.23**	83.33	90.47	90.47
KAB-BTN	90.24	87.80	90.24	**92.68**	90.24	87.80	**92.68**	**92.68**	80.48	85.36	85.36
KAB-DFL	65.38	63.46	65.38	63.46	65.38	63.46	63.46	**96.15**	80.76	**96.15**	**96.15**
JLF-TNS	50.00	52.50	50.00	47.50	50.00	52.50	50.00	75.00	70.00	75.00	72.50
JLF-CST	74.19	64.51	74.19	64.51	74.19	61.29	67.74	67.74	67.74	70.96	70.96
JLF-BTN	79.31	86.20	79.31	**93.10**	79.31	86.20	**93.10**	**93.10**	65.51	89.65	89.65
JLF-DFL	52.50	45.00	52.50	42.50	52.50	42.50	42.50	57.50	70.00	55.00	55.00
TNS-CST	58.06	51.61	67.74	74.19	67.74	51.61	70.96	64.51	67.74	67.74	67.74
TNS-BTN	72.41	79.31	62.06	82.75	72.41	86.20	75.86	**100.00**	79.31	86.20	86.20
TNS-DFL	42.50	52.50	42.50	52.50	42.50	52.50	52.50	40.00	42.50	37.50	40.00
CST-BTN	55.00	55.00	55.00	55.00	65.00	60.00	60.00	95.00	**100.00**	95.00	95.00
CST-DFL	58.06	48.38	58.06	70.96	58.06	54.83	70.96	51.61	54.83	54.83	54.83
BTN-DFL	72.41	75.86	72.41	93.10	72.41	86.20	93.10	**96.55**	79.31	93.10	93.10
Multiclass	31.12	24.50	11.92	28.47	17.88	25.16	25.16	41.05	25.82	33.11	38.41

Table 5: Word vs Sentence Level Dialect Identification Applied to Algerian Dialect Corpus

6. Conclusions And Future Work

In this work, we presented a word-level and sentence-level based dialect identification system. At word level, we introduced mV and MV rules, and showed how mV performed well for Algerian dialects. However, at the sentence level, the binary classification model provides better scores for Arabic dialects particularly. Indeed, the average accuracy for the five Arabic dialects (Algerian, Tunisian, Moroccan, Syrian and Palestinian) is over 92%. The average accuracy for Algerian dialects is ranging between 62%, - at word level - to 76% - at sentence level -.
For future work, we aim to cover more dialects with significant sizes, especially for Algerian which is considered as an under-resourced language since it is less written and only few texts are present on the web.

References

[1] Bougrine, S., Cherroun, H., & Abdelali, A. (2018, April). Spoken Arabic Algerian Dialect Identification. In *IEEE 2nd International Conference on Natural Language and Speech Processing (ICNLSP 2018)*, Algeria, (pp.1-6). IEEE.

[2] Moftah, M., Fakhr, M. W., & El Ramly, S. (2018, April). Arabic Dialect Identification based on Motif Discovery using GMM-UBM with Different Motif Lengths. In *IEEE 2nd International Conference on Natural Language and Speech Processing (ICNLSP 2018)* (pp. 1-6). IEEE.

[3] Alshutayri, A., & Albarhamtoshy, H. (2011, November). Arabic Spoken Language Identification System (ASLIS): A Proposed System to Identifying Modern Standard Arabic (MSA) and Egyptian Dialect. In *International Conference on Informatics Engineering and Information Science* (pp.375-385). Springer, Berlin, Heidelberg.

[4] Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., ... & Renals, S. (2015). Automatic Dialect Detection in Arabic Broadcast Speech. *arXiv preprint* arXiv:1509.06928.

[5] Khurana, S., Najafian, M., Ali, A., Al Hanai, T., Belinkov, Y., & Glass, J. (2017). QMDIS: QCRI-MIT Advanced Dialect Identification System. In Proc. *Interspeech 2017* (pp.2591-2595).

[6] Zaidan, O. F., & Callison-Burch, C. (2014). Arabic Dialect Identification. *Computational Linguistics*, **40**(1), 171-202.

[7] Harrat, S., Meftouh, K., Abbas, M., & Smaili, K. (2014). Building Resources for Algerian Arabic Dialects. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[8] Alshutayri, A., & Atwell, E. (2018, May). Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools* (p. 54).

[9] Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaili, K. (2015, October). Machine Translation Experiments on PADIC: A parallel Arabic Dialect Corpus. In *The 29th Pacific Asia conference on language, information and computation*.

[10] Guellil, I., Azouaou, F., Abbas, M., & Sadat, F. (2017, May). 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* (colocated with EAMT2017).

[11] Harrat, S., Abbas, M., Meftouh, K., & Smaili, K. (2013, August). Diacritics Restoration for Arabic Dialect Texts. In *INTERSPEECH* (pp. 1429-1433).

[12] Eldesouki, M., Samih, Y., Abdelali, A., Attia, M., Mubarak, H., Darwish, K., & Laura, K. (2017). Arabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM. *arXiv preprint* arXiv:1708.05891.

[13] Samih, Y., Eldesouki, M., Attia, M., Darwish, K., Abdelali, A., Mubarak, H., & Kallmeyer, L. (2017). Learning from Relatives: Unified Dialectal Arabic Segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (CoNLL 2017) (pp. 432-441).

[14] Elfardy, H., & Diab, M. (2013). Sentence Level Dialect Identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Vol. **2**, pp. 456-461).

[15] Sadat, F., Kazemi, F., & Farzindar, A. (2014, July). Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the first international workshop on Social media retrieval and analysis* (pp. 35-40). ACM.

[16] Durandin, O. V., Strebkov, D. Y., & Hilal, N. R. (2016). Automatic Arabic Dialect Classification. In *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue"*(2016) (pp. 1-13).

[17] Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaili, K. (2015, April). Cross-dialectal Arabic Processing. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 620-632). Springer.

[18] Tillmann, C., Mansour, S., & Al-Onaizan, Y. (2014). Improved Sentence-level Arabic Dialect Classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects* (pp. 110-119).

[19] Malmasi, S., & Dras, M. (2015). Language Identification Using Classifier Ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects* (pp. 35-43).

[20] Al-Badrashiny, M., Elfardy, H., & Diab, M. (2015). Aida2: A Hybrid Approach for Token and Sentence Level Dialect Identification in Arabic. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (pp. 42-51).

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. *Journal of machine learning research*, 12(Oct), 2825-2830.