SQU-CS @ NADI 2022: Dialectal Arabic Identification using One-vs-One Classification with TF-IDF Weights Computed on Character n-grams

Abdulrahman AAlAbdulsalam
Department of Computer Science
Sultan Qaboos University, Oman
a.aalabdulsalam@squ.edu.om

Abstract

In this paper, I present an approach using one-vs-one classification scheme with TF-IDF term weighting on character n-grams for identifying Arabic dialects used in social media. The scheme was evaluated in the context of the third Nuanced Arabic Dialect Identification (NADI 2022) shared task for identifying Arabic dialects used in Twitter messages. The approach was implemented with logistic regression loss and trained using stochastic gradient decent (SGD) algorithm. This simple method achieved a macro F1 score of 22.89% and 10.83% on TEST A and TEST B, respectively, in comparison to an approach based on AraBERT pretrained transformer model which achieved a macro F1 score of 30.01% and 14.84%, respectively. My submission based on AraBERT scored a macro F1 average of 22.42% and was ranked 10 out of the 19 teams who participated in the task.

1 Introduction

Arabic is well known for its rich morphology and complex system of inflectional forms (Habash et al., 2005). While a word in English may have few inflections, a word in Arabic contains many more inflectional forms depending on tense, number, person, mood, gender and voice (Neme and Laporte, 2013). Arabs mostly communicate informally using a continuum of dialects that vary from the east in the Arabian peninsula to the west in the North African region. These dialects add another layer of complexity since they differ at the phonological, morphological, lexical and syntactic levels (Abdul-Mageed et al., 2018). Despite dialects are predominantly used in spoken form, heavy usage of the written form is becoming very popular especially in social media platforms (Mubarak and Darwish, 2014).

Most of past research on Arabic Natural Language Processing (ANLP) have mainly focused on Modern Standard Arabic (MSA), the formal language of communication used in the Arab world. However, recently, Arabic dialects have gained more attention by researchers especially the Egyptian dialect (Guellil et al., 2021). Research on Arabic dialects involved improving parts-of-speech tagging (Alharbi et al., 2018), named entity recognition (Zirikly and Diab, 2015), parsing & grammar (Albogamy et al., 2017) and machine translation (Harrat et al., 2019). One key finding is that higher-level language tasks on Arabic dialects benefit substantially from the application of low-level pre-processing techniques that focus on better segmentation and word morphology analysis (El Kah and Zeroual, 2021; Duwairi and El-Orfali, 2014).

Arabic is considered a low-resource language when compared to other languages (Sajjad et al., 2020). This makes it challenging to utilize pre-existing approaches based on supervised machine learning (El Mekki et al., 2020). Recent works have focused on the use of few-shot or zero-shot learning techniques for Arabic dialects with promising results (Khalifa et al., 2021b,a).

The subtask of identifying Arabic dialect at the country-level was conducted as part of the third Nuanced Arabic Dialect Identification shared task: NADI 2022 (Abdul-Mageed et al., 2022). Similar subtask was organized in prior years during NADI 2020 (Abdul-Mageed et al., 2020) and NADI 2021 (Abdul-Mageed et al., 2021) shared tasks. Past attempts used a variety of approaches ranging from classical machine learning, to ensemble-based classification, and deep learning mutli-task transformer-based neural networks. The best performing methods reported for this subtask utilized the transformer-based models trained with multi-task prediction (Abdul-Mageed et al., 2021).

The current paper describes an approach based on one-vs-one classifiers trained with TF-IDF term weights on character n-grams for identifying Arabic dialects. The motivation for this approach is

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 436 - 441

the success of recent methods that exploit subword units for learning as opposed to the individual word tokens (Baniata et al., 2021; Alyafeai et al., 2022). The subword units representation work better in practice especially for Arabic and help reduce out-of-vocabulary (OOV) tokens; a common problem in natural language processing tasks. The proposed approach can be used as a baseline performance on the task of Arabic dialect identification.

2 Data

Shared task organizers have prepared and distributed two datasets with country-level labels for the task participants that can be utilized for system development as shown in Table 1. Each sample in

Country Label	TRAIN	DEV
egypt	4283	1041
iraq	2729	664
ksa	2140	520
algeria	1809	430
oman	1501	355
syria	1287	278
libya	1286	314
tunisia	859	173
morocco	858	207
lebanon	644	157
uae	642	157
jordan	429	104
kuwait	429	105
yemen	429	105
palestine	428	104
bahrain	215	52
qatar	215	52
sudan	215	53
TOTAL	20398	4871

Table 1: Country-level label counts for the shared task TRAIN and DEV datasets.

the datasets is a single tweet message containing the original text with user mentions and website links replaced with the 'USER' and 'URL' tokens, respectively. In addition, two datasets TEST-A (4,758 samples) and TEST-B (1,474 samples) were distributed without labels and were used for final evaluation of participating teams submissions.

3 System Description

3.1 Data Preprocessing

The text in the datasets was preprocessed as follows:

- Remove all non-Arabic printable ASCII characters (hexdecimal codes `21` to `7E`).
- Remove any Arabic diacritic marks (Unicode ranges `0617-061A` and `064B-0652`).
- Normalize by replacing three or more repetitions of the same letter with two occurrences. For instance, the word مرررحباااا will be normalized to مرحباا in which many repetitions of the letters ا and ر were reduced to two occurrences only.
- Normalize by replacing variants of the letter Alif أ إ آ with the letter ا, the letter ة with the letter ه, and the letter ى with letter ي.

3.2 Algorithms & Implementations

Two submissions were sent to the task organizers for evaluation. First submission (henceforth will be referred to as `OVO-LR`) used one-vs-one binary classifiers implementing the Logistic Regression (LR) loss function defined in the equation below and trained with stochastic gradient descent (SGD) algorithm.

$$\sum_i (-y_i \log(h) \; + \; 1 - y_i \log(1-h)) \qquad (1)$$

Where h is the predicted probability of the true class label obtained with the sigmoid function ($\sigma(z) = \frac{1}{1-e^{(-z)}}$) and y_i is the true binary label (0 or 1).

A vocabulary consisting of character n-grams where $2 \leq n \leq 5$ (see table 2 for an example) was generated from all the text in TRAIN, DEV and TEST sets for the shared task. If the length of a particular word in the input text is less than 2 or greater than 5 then it was appended to the vocabulary. The Term Frequency-Inverted Document Frequency (TF-IDF) weights were computed for each word in the resulting vocabulary where each sample (a single Twitter message) in the dataset is considered a document. Each sample in the input was represented as a vector of TF-IDF weights using the one-hot enconding scheme. Eventually a collection of input samples in a dataset were presented for the classifier in a sparse matrix format containing the "bag of character n-grams" as input features.

The one-vs-one LR classifiers were trained using SGD algorithm with the L2 regularization penalty. Due to the skewed distribution of the class labels in both TRAIN and DEV sets, the class weights were set to be inversely proportional to label counts distribution found in the training data.

input text	character n-grams
الطرحه في النفر جنيه	ال لط طر رح حه الط لطر طرح رحه الطر لطرح طرحه الطرح لطرحه في ال لن نف فر الن لنف نفر النف لنفر النفر جن ني يه جني نيه جنيه

Table 2: Sample input text converted into character n-gram representation ($2 \leq n \leq 5$).

This submission was implemented using the following python `scikit-learn` packages using the default settings for other parameters:

- `SGDClassifier(loss='log', class_weights='balanced')`
- `OneVsOneClassifier()`
- `TfIdfVectorizer(analyzer='char', ngram_range=(2, 5))`

The second submission (`AraBERT-NADI`) used pretrained AraBERT transformer model (`bert-base-arabertv02-twitter`) which was trained on 60 millions tweets containing various Arabic dialects (Antoun et al., 2020). The model was adapted for the dialect identification task by retraining the prediction layer using the TRAIN set only for 10 epochs (learning rate = 2×10^{-5}, adam epsilon = 1×10^{-8}, training batch size = 16).

4 Results

Table 3 lists the results obtained for my official submissions on the DEV, TEST A and TEST B sets including the best scores for TEST A and TEST B in the task. The official metric used for ranking submissions in the task is the macro-averaged F1 scores. The final official scores show that `AraBERT-NADI` achieved better F1 score with +7.12% higher than `OVO-LR` for TEST A. The best overall F1 score for TEST A in the task is +6.47% higher than `AraBERT-NADI`. In addition, `AraBERT-NADI` scored +4.01% percentage points higher than `OVO-LR` for TEST B. The best F1 score obtained for TEST B is +4.11% percentage points higher than `AraBERT-NADI`. The reason for the sharp difference in performance between TEST A and TEST B could be explained by the fact that TEST B only contains a subset of the total 18 country labels in TEST A[1]. Another possible explanation is possible mismatch of label distribution between the training data (TRAIN and DEV) and TEST B. This will affect the performance of classification models which were trained to place significant weight on feature terms for the majority class labels and, therefore, become biased towards making positive predictions for the majority class label (Padurariu and Breaban, 2019). The difference in label distributions could justify the drastic drop in performance obtained in TEST B set in comparison to TEST A set (-12.06%, -15.17% and -17.53% points drop in F1 scores for `OVO-LR`, `AraBERT-NADI` and `BEST*`, respectively).

Table 4 lists the per-country label breakdown of the scores obtained on the DEV set for the submitted models. Overall, the `AraBERT-NADI` model performed better on most of country labels than the `OVO-LR` classifier. Both models performed worse on country labels with low distribution in the training data especially for the GULF dialects: Bahrain, Qatar and Yemen. An exception to this is the Arabic dialect of Sudan in which both models performed in-par or better than other dialects with much more training samples (e.g., Omani dialect). This maybe due to the fact that Sudanese dialect contain unique phrases not shared by many other Arabic dialects (see table 5). The `OVO-LR` scored better on Qatari and Kuwaiti dialects than `AraBERT-NADI` classifier. This may be because `OVO-LR` model was trained to increase the weight of low distribution class labels (i.e., assign more weight to samples from lower represented class labels). Both models obtained zero score on Bahraini dialect which is spoken in the GULF region. After manually inspecting the samples of Bahraini dialect in the TRAIN and DEV sets, it is clearly that there is a major difference in discourses between the two sets. Most of the samples in the TRAIN set include topics of sports genre and predominantly contain masculine pronouns. On the other hand, most of samples in the DEV set include topics of social genre with predominantly feminine pronouns.

Table 5 shows the top n-gram features used by `OVO-LR` model to classify each dialect. Many features are shared across dialects especially bi-grams such as شو, هه, وش and في. Notable discriminating features are n-grams that indicate country names such as تونس for Tunisia, لبنان for Lebanon, عراق for Iraq, and لغ for Morocco. Country names are not good features for identifying dialects per se, which indicate one of the limitations of bag of words ap-

[1] According to the task organizers, TEST-B covers k country-level dialects, where k is unknown.

Submission	Dataset	Acc.	Rec.	Prec.	Macro-F1
OVO-LR	TEST A	36.34	22.97	23.18	22.89
	TEST B	20.69	11.18	15.03	10.83
	DEV	35.04	22.99	22.29	21.89
AraBERT-NADI	TEST A	46.85	29.75	34.57	30.01
	TEST B	30.12	16.80	21.32	14.84
	DEV	47.32	29.12	34.57	29.16
BEST*	TEST A	53.05	35.22	41.89	36.48
	TEST B	36.97	20.48	25.82	18.95

Table 3: Official task submissions results; BEST* are the top scores obtained in the task.

Country label	AraBERT-NADI			OVO-LR		
	Prec.	Rec.	F1	Prec.	Rec.	F1
algeria	63.21	41.16	49.86	42.89	43.49	43.19
bahrain	0.00	0.00	0.00	0.00	0.00	0.00
egypt	62.41	89.15	73.42	60.30	66.09	63.06
iraq	61.55	56.17	58.74	47.84	46.69	47.26
jordan	33.33	6.73	11.20	9.35	12.50	10.70
ksa	36.17	55.58	43.82	34.25	21.54	26.45
kuwait	20.59	6.67	10.07	10.16	18.10	13.01
lebanon	44.44	12.74	19.80	16.06	14.01	14.97
libya	47.26	43.95	45.54	39.66	29.94	34.12
morocco	46.25	17.96	25.87	19.01	11.17	14.07
oman	25.71	33.24	28.99	27.87	19.15	22.70
palestine	20.33	24.04	22.03	13.45	15.38	14.35
qatar	0.00	0.00	0.00	3.49	5.77	4.35
sudan	33.33	52.83	40.88	16.38	35.85	22.49
syria	24.51	22.66	23.55	19.77	12.59	15.38
tunisia	36.11	22.54	27.76	19.62	23.70	21.47
uae	27.06	29.30	28.13	16.10	33.12	21.67
yemen	40.00	9.52	15.38	5.15	4.76	4.95

Table 4: Breakdown of scores obtained on the DEV set for each country label in the dataset.

proach used in OVO-LR and the nature of the data used in the task.

5 Conclusion

In this paper, I presented my attempt to identify country-level Arabic dialects used in Twitter messages. The approach based on simple one-vs-one classifiers using logistic regression loss showed good baseline performance on the testing sets for the shared task in comparison to BERT-based transformer model (AraBERT) that was pretrained on 60 million Arabic tweets.

Acknowledgements

I would like to thank NADI 2022 shared task organizers for their efforts in preparing Arabic dialect datasets and releasing them to the research community. Such resources facilitate objective comparison of different methods and advance our knowledge of state-of-the-art in the field of Arabic Natural language Processing.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110,

algeria	هذ نت صح في ربي تاع ربي واش ران كي اك اش را
bahrain	الأ أن إت الأ حاد طو هه أبو وش تحاد إل وش لأ
egypt	ربنا ان بت كد دي بقي هت بق يا ده مش
iraq	يم عر نو ريد نه كول حجي مو ار كو حج بچ هاي اي اني عراق اكو
jordan	حدا حلو هيك حكي اشي رب له ابو اي سطي بد له زم طب دا نور بح
ksa	ره ون كث ما عط يال كذا مو لك سوي صي يال دري اب در ها لك ني الا وش ذا
kuwait	بخ مو لجو الجو حين لج ابي يت اق بعد بي يال حي حين هلا ني هل لحين حي ما نج صج
lebanon	ساح انو هي رح عم يدي ما كر حق هيد دا يل نان ار رح يا لبنا هيك عم لبنان شو يا
libya	ات هك ليب ليبي نخ هلب خوي هكي ربي هذ توا في نق دير شن
morocco	هه داك لمغ ال غا لي فال اد ديال شي هاد
oman	ابا بع شي ما يد عاد سوي تو هاا ترا بعد هيه بو هال
palestine	بك لله حدا سي هد يسع ير حد يا شو هيك ها يسعد بت اشي مش
qatar	لج جي سنك لى نك واو لق عسى لج كم مب ما
sudan	بي ساي شن يهو لي شنو نو دي ياخ دا
syria	يكي عال بدي ما كت لك شو نو بع بدي عم هال اي انو
tunisia	با تونس تش اش كا وش في كان انت قال تو موش في
uae	حظ وي ول بي بك شو يت به تب حبه الي لب ني غل تك يال تك
yemen	لج سقط نك ردي يمن سقط جوو بت حبيش بيش لس ايش زول لي جو

Table 5: Selected subset of top character n-gram features used by the OVO-LR model to classify Arabic dialects.

Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim A. Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced arabic dialect identification shared task. *CoRR*, abs/2103.08466.

Fahad Albogamy, Allan Ramsay, and Hanady Ahmed. 2017. Arabic tweets treebanking and parsing: A bootstrapping approach. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 94–99.

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2018. Part-of-speech tagging for arabic gulf dialect using bi-lstm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zaid Alyafeai, Maged S Al-shaibani, Mustafa Ghaleb, and Irfan Ahmad. 2022. Evaluating various tokenizers for arabic text classification. *Neural Processing Letters*, pages 1–23.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Laith H Baniata, Isaac KE Ampomah, and Seyoung Park. 2021. A transformer-based neural machine translation model for arabic dialects that utilizes subword units. *Sensors*, 21(19):6509.

Rehab Duwairi and Mahmoud El-Orfali. 2014. A study of the effects of preprocessing strategies on sentiment analysis for arabic text. *Journal of Information Science*, 40(4):501–513.

Anoual El Kah and Imad Zeroual. 2021. The effects of pre-processing techniques on arabic text classification. *Int. J*, 10(1):1–12.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of bert and n-gram features for nuanced arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Nizar Habash, Owen Rambow, and George Anton Kiraz. 2005. Morphological analysis and generation for arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

Muhammad Khalifa, Muhammad Abdul-Mageed, and Khaled Shaalan. 2021a. Self-training pre-trained language models for zero-and few-shot multi-dialectal arabic sequence labeling. *arXiv preprint arXiv:2101.04758*.

Muhammad Khalifa, Hesham Hassan, and Aly Fahmy. 2021b. Zero-resource multi-dialectal arabic natural language understanding. *arXiv preprint arXiv:2104.06591*.

Hamdy Mubarak and Kareem Darwish. 2014. Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Alexis Amid Neme and Eric Laporte. 2013. Pattern-and-root inflectional morphology: the arabic broken plural. *Language Sciences*, 40:221–250.

Cristian Padurariu and Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 176–185.