iCompass Working Notes for the Nuanced Arabic Dialect Identification Shared task.

Abir Messaoudi **Chayma Fourati** **Hatem Haddad** **Moez Ben HajHmida**
iCompass, Tunisia
{abir,chayma,hatem,moez}@icompass.digital

Abstract

We describe our submitted system to the Nuanced Arabic Dialect Identification (NADI) shared task. We tackled only the first subtask (Subtask 1). We used state-of-the-art Deep Learning models and pre-trained contextualized text representation models that we fine-tuned according to the downstream task in hand. As a first approach, we used BERT Arabic variants: MARBERT with its two versions MARBERT v1 and MARBERT v2, then we combined MARBERT embeddings with a CNN classifier, and finally, we tested the Quasi-Recurrent Neural Networks (QRNN) model. The results found show that version 2 of MARBERT outperforms all of the previously mentioned models on Subtask 1.

1 Introduction

Nowadays, social media is spread all over Arabic countries where people tend to express themselves in their own local dialect. Since it has different variants and dialects across the world, Arabic dialect identification presents a challenging task. Even if some dialects share some vocabulary, they still differ according to countries, where each dialect has its own specifications. Because of the massive amount of such content, automatic identification of Arabic dialects becomes crucial. Following the first (Abdul-Mageed et al., 2020b) and second (Abdul-Mageed et al., 2021) Nuanced Arabic Dialect Identification (NADI 2020 and NADI 2021), NADI 2022 subtask 1 focuses on identifying the Arabic dialect of a given text, especially on social media sources where there is no established standard orthography like Modern Standard Arabic (MSA) (Abdul-Mageed et al., 2022). The first attempts to tackle this challenge identified different Arabic dialects categories in addition to MSA: Maghrebi, Egyptian, Levantine, Gulf, and Iraqi (Zaidan and Callison-Burch, 2011). In (El-Haj et al., 2018) authors proposed 4 Arabic dialects categories by merging the Iraqi with the Gulf.

The paper is structured as follows: Section 2 provides a concise description of the used dataset, its statistics, and pre-processing techniques. Section 3 describes the used systems and the experimental setup to build models for Country-level dialect identification. Section 4 presents and discusses the obtained results. Finally, section 5 concludes and points to possible directions for future work.

2 Data Description

The provided training dataset of the competition (Abdul-Mageed et al., 2022) dedicated for the first subtask consists of around 25k tweets written in eighteen Arabic dialects including: Egypt, Iraq, KSA, Algeria, Oman, Syria, Libya, Tunisia, Morocco, Lebanon, UAE, Jordan, Kuwait, Yemen, Palestine, Bahrain, Qatar, and Sudan. Figure 1 presents the distribution of the tweets over the eighteen labels. In fact, the training dataset is imbalanced and presents skewed class proportions. We notice the domination of Egypt and Iraq tweets compared to the other countries.

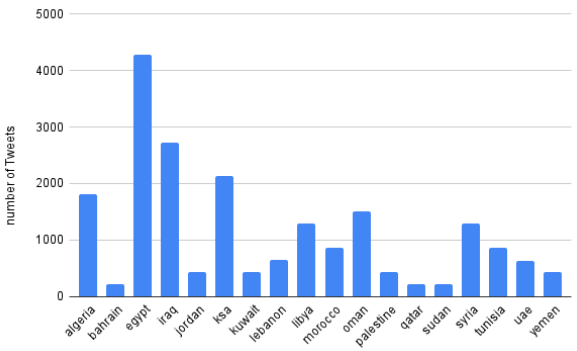

Figure 1: The distribution of tweets according to the 18 classes.

2.1 Data pre-processing

In order to normalize the dataset, we managed to do several strategies of cleaning. In fact, we remove

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 415 - 419

all non Arabic tokens, including ones like USER, URL, < LF >. Emojis were also removed. We normalize all the hashtags by simply decomposing them and we ended by removing successive white spaces.

In order to validate our models, we use the training and development datasets provided by NADI 2022 competition. Table 1 presents statistics of the training and development datasets for Subtask 1.

Data	# Sentences
Training	20398
Development	4871

Table 1: Training and development datasets statistics for Subtask 1.

3 System Description

Different deep learning architectures and pre-trained language models were used in order to achieve the best results.

3.1 MARBERT

MARBERT, also by (Abdul-Mageed et al., 2020a) is a large-scale pretrained language model using BERT base's architecture and focusing on the various Arabic dialects. It was trained on 128 GB of Arabic tweets. The authors chose to keep the tweets that have at least 3 Arabic words. Therefore, tweets that have 3 or more Arabic tokens without removing non-Arabic (foreign languages) ones (15.6 billion Arabic and non-Arabic tokens). This is because dialects are often times mixed with other foreign languages. MARBERT enhances the language variety as it focuses on representing the previously underrepresented dialects and Arabic variants. MARBERT v2 is the second version of MARBERT pre-trained on the same MSA data as ARBERT in addition to AraNews dataset but with a bigger sequence length of 512 tokens for 40 epochs.

3.2 Convolutional Neural Network

The dataset was tokenized using both versions of MARBERT (v1 and v2) tokenizer, mapping words to their indexes. MARBERT embedding matrix was used at the embedding layer level. Then, Convolutional Neural Network (CNN) model was used as classifier and a fully connected layer with a softmax activation function in order to predict label's probabilities with the following hyper-parameters: batch size of 32, max sequence length of 64, and 4 epochs.

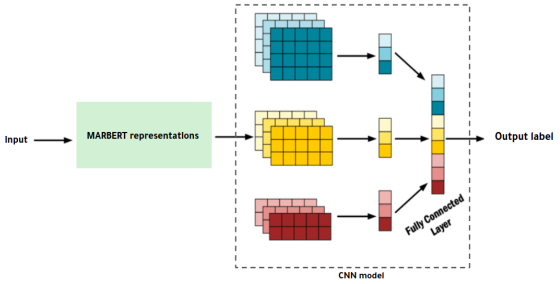

Figure 2: MARBERT + CNN architecture.

3.3 Quasi-recurrent Neural Network

Quasi-recurrent neural network (QRNN) (Bradbury et al., 2016) represents an architecture that combines the sequential manner of treating the input tokens from Recurrent Neural Networks (RNNs) and the parallel processing fashion of Convolutional Neural Networks (CNNs) to allow a longer term dependency window while also addressing several issues faced when using both architectures separately. Stacked QRNNs are reported to have a better predictive accuracy than stacked LSTMs of the same hidden size (Bradbury et al., 2016). MARBERT v2 was used as the embedding layer, followed by the QRNN model. Hyper-parameters used are: batch size of 32, max sequence length of 64, and 8 epochs.

Figure 3 represents details of the QRNN architecture.

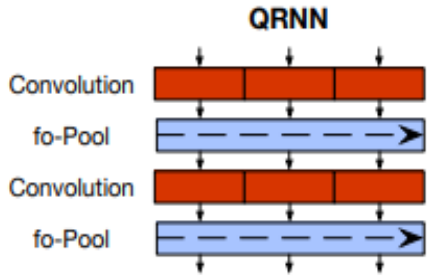

Figure 3: QRNN architecture.(Bradbury et al., 2016)

3.4 System submission

As an approach, we used the Arabic BERT (Devlin et al., 2019) variant MARBERT (Abdul-Mageed et al., 2020a) (second version) since it was trained mostly on dialectal Arabic which was underrepresented in previous pretrained models. Since this

task's data is multi-dialectal, this model is expected to achieve the best performance. We used the training dataset provided by the NADI 2022 shared task that covers 18 dialects (total of 20K tweets, the same as NADI 2021) (Abdul-Mageed et al., 2022). We trained our model on a Google Cloud GPU of 8 cores using Google Colaboratory. The final model hyper-parameters that we used to make the submission are:

- Model name: MARBERT v2
- Number of epochs: 4
- Learning rate: 2e-5
- Batch size: 32
- Max sequence length: 64

4 Results and Discussion

We submitted one run to subtask 1: trained on the provided training dataset. This subtask is a multi-class classification problem, including eighteen labels.

Model	Macro-F1	Accuracy
MARBERT v1 + CNN	0.12	0.39
MARBERT v2 + CNN	0.14	0.40
MARBERT v2 + QRNN	0.26	0.41
MARBERT v2	**0.33**	**0.50**

Table 2: Results of different models on the development dataset.

Table 2 presents the results of experiments performed for this subtask. Preliminary results on the development dataset showed that a fine-tuned MARBERT v2 achieved the best performances compared to the other three models in term of Accuracy and marco-F1.

Using MARBERT v2 as the embedding layer followed by the QRNN outperforms MARBERT v2 as the embedding layer followed CNN. Fine-tuning the pre-trained model MARBERT with QRNN looks very promising for small sized annotated Arabic dialects data as mentioned in (Bennessir et al., 2022) but further experiments are needed to substantiate this assumption.

We notice that the data imbalance decreased the model performance in terms of macro-F1. Figures 4 and 5 show confusion matrices where the classes most correctly classified are: 2 for Egypt, 3 for Iraq and 5 for KSA, which are the countries with higher presence in the training dataset. The model trained with MARBERT + CNN architecture, in Figures 4, tends to always predict the oversampled classes, which explains the low Macro-F1 score. In fact, most of Omanian (10), Syrian (17) and Bahrainian (16) sentences are predicted as Saudian (5). Most of Moroccan (9) sentences are predicted as Algerians (0).

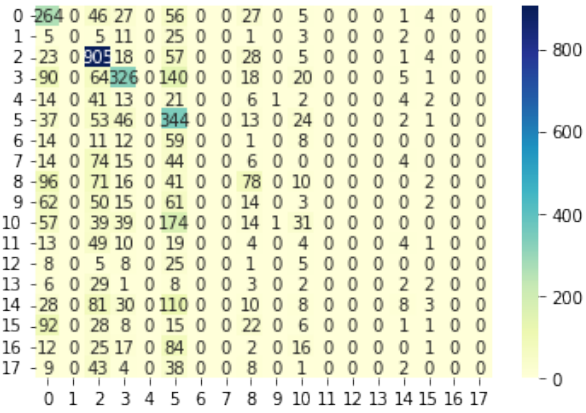

Figure 4: Confusion matrix of the MARBERT v2 + CNN model. ('0:alg','1:bah', '2:egy', '3:irq', '4:jor', '5:ksa', '6:kuw', '7:leb', '8:lib', '9:mor' ,'10:om', '11:pal', '12:qatar' ,'13:sud', '14:syr', '15:tun', '16:uae', '17:yem')

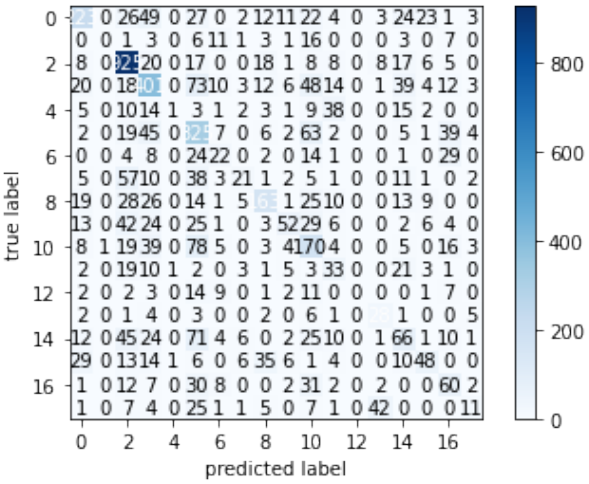

Figure 5: Confusion matrix of the MARBERT v2 model. ('0:alg','1:bah', '2:egy', '3:irq', '4:jor', '5:ksa', '6:kuw', '7:leb', '8:lib', '9:mor' ,'10:om', '11:pal', '12:qatar' ,'13:sud', '14:syr', '15:tun', '16:uae', '17:yem')

4.1 Official submission results

NADI provides two test sets: Test-A and Test-B. TEST-A covers 18 country-level dialects, containing 4,758 tweets, whereas the second test set

(TEST-B) covers an unknown country-level dialects. Then, the subtask score is calculated using the average score between the two test sets. Tables 3, 4 and 5 review the official results of iCompass system for NADI (resp. Test-A and Test-B) on the test dataset against the top three ranked systems.

Team	Rank	Macro-F1	Accuracy
rematchka	1	36.4807	53.0475
GOF	2	35.6825	52.1017
UniManc	3	34.7780	52.3329
iCompass	**4**	**33.7000**	**51.9126**

Table 3: Leaderboard of Test-A of Subtask 1.

Team	Rank	Macro-F1	Accuracy
UniManc	1	18.9481	36.8385
mtu_fiz	2	17.6715	33.9213
rematchka	3	17.6361	36.49936
iCompass	**7**	**16.937**	**34.9389**

Table 4: Leaderboard of Test-B of Subtask 1.

Team	Rank	Average Macro-F1
rematchka	1	27.06
UniManc	2	26.86
GOF	3	26.44
iCompass	**5**	**25.32**

Table 5: Leaderboard of Subtask 1.

5 Conclusion

In this work, MARBERT (Abdul-Mageed et al., 2020a) in its second version was used to identify Country-level dialect. The best results were obtained by MARBERT v2 with specific hyperparameters, which was selected for the final submission. Future work would involve building a multi-script Arabic dialects language model including Arabic script and Latin script based characters. Taking as example, Tunisians, who tend to express themselves using an informal way called TUNIZI (Fourati et al., 2021) that represents the Tunisian text written using Latin characters and numbers instead of Arabic letters.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim 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, AbdelRahim 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)*.

Mohamed Aziz Bennessir, Malek Rhouma, Hatem Haddad, and Chayma Fourati. 2022. icompass at arabic hate speech 2022: Detect hate speech using qrnn and transformers. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 176–180.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2016. Quasi-recurrent neural networks. *arXiv preprint arXiv:1611.01576*.

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.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large Tunisian Arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset

of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.