NLP_DI at NADI 2024 shared task: Multi-label Arabic Dialect Classifications with an Unsupervised Cross-Encoder

Vani Kanjirangat
IDSIA-USI/SUPSI, Switzerland
vanik@idsia.ch

Tanja Samardžić
URPP Language and Space, UZH
tanja.samardzic@uzh.ch

Ljiljana Dolamic
armasuisse S+T, Switzerland
Ljiljana.Dolamic@armasuisse.ch

Fabio Rinaldi
IDSIA-USI/SUPSI, Switzerland
fabio.rinaldi@idsia.ch

Abstract

We report the approaches submitted to the NADI 2024 Subtask 1: Multi-label country-level Dialect Identification (MLDID). The core part was to adapt the information from multi-class data for a multi-label dialect classification task. We experimented with supervised and unsupervised strategies to tackle the task in this challenging setting. Under the supervised setup, we used the model trained using NADI 2023 data and devised approaches to convert the multi-class predictions to multi-label by using information from the confusion matrix or calibrated probabilities. Under unsupervised settings, we used the Arabic-based sentence encoders and multilingual cross-encoders to retrieve similar samples from the training set, considering each test input as a query. The associated labels are then assigned to the input query. We also tried variations, such as co-occurring dialects derived from the provided development set. We obtained the best validation performance of 48.5% F-score using one of the variations with an unsupervised approach and the same approach yielded the best test result of 43.27% (Ranked 2).

1 Introduction

Arabic is a language spoken by a large community of about 400 million people, which is widely distributed around different countries and regions. Modern Standard Arabic (MSA), the official language in many Arabic-speaking countries, differs from the regional varieties lexically, syntactically, and phonetically (Zaidan and Callison-Burch, 2014).

Arabic Dialect Identification (ADI) deals with identifying the dialect of a given Arabic input utterance. Some of the most popular datasets in ADI include: The ADI VarDial dataset (Zampieri et al., 2017, 2018), which includes Arabic text that is both speech transcribed and transliterated (Malmasi et al., 2016; Ali et al., 2016), Arabic Online Commentary (AOC), which includes a large-scale repository of Arabic dialects obtained from reader commentary of online Arabic newspapers (Zaidan and Callison-Burch, 2011), Multi Arabic Dialect Applications and Resources (MADAR) corpus constitutes parallel sentences written in different Arabic city dialects from travel domain (Bouamor et al., 2019) etc. The NADI shared task started in 2020 and presented continued efforts in the ADI, including country-wise, province-wise, and region-wise dialect identification tasks (Abdul-Mageed et al., 2020, 2021, 2022, 2024). The main limitation of these aforementioned datasets is that they are mono-labeled, which means they are multi-class, where each input sample belongs exactly to one of the classes/dialects. The 2024 NADI ADI task focuses on multi-label dialect identifications, meaning an input sample can belong to more than one dialect class. In general, the approaches for dialect classifications ranged from n-gram and machine learning approaches (Touileb, 2020; Younes et al., 2020; AlShenaifi and Azmi, 2020; Harrat et al., 2019; Çöltekin et al., 2018; Butnaru and Ionescu, 2018) to ensembles El Mekki et al. (2020) and pre-trained neural models (AlKhamissi et al., 2021; El Mekki et al., 2021; Elaraby and Abdul-Mageed, 2018; Ali, 2018).

In this paper, we describe our solutions submitted to the NADI shared task 2024 (Abdul-Mageed et al., 2024), subtask-1, which targets the challenging task of Multi-label country-level Dialect Identification (MLDID). The unavailability of multi-label training data poses a major roadblock in this task. Hence, the main challenge is to adapt the available multi-class training data for multi-label predictions.

This paper is organized as follows: The data statistics are described in Section 2, methods used are discussed in Section 3, experimental results are reported in Section 4, followed by conclusions in Section 5.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 742–747

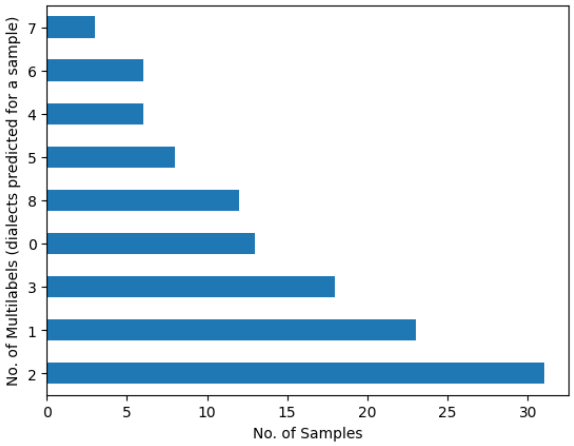

Figure 1: Number of samples to the number of dialects belonging to each sample

2 Data

Subtask 1 of NADI 2024 provided a development set with 8 country dialects constituting 100 samples. The samples were multi-labeled, i.e., each sample can belong to more than one country dialect. The label distribution is shown in Figure 1, which depicts that we have 13 samples that did not belong to any of the 8 dialects and about 12 samples belonging to all of the 8 dialects. Most of the samples were belonging to at least two classes.

The test set included 1000 samples, which are supposed to be multi-dialectal, while the number of dialects was kept unknown, with a maximum number of dialects specified as 18. Hence, the number of multi-label classes,*l*, for a single instance can be *l<=18*.

3 Models and Methods

This section describes the techniques used for NADI-2024 MLDID SubTask 1. The task at hand considers an input utterance that can belong to multiple dialect classes, which is inherently challenging. The challenge becomes multi-faceted by factors such as the unavailability of multi-label training data and the unknown number of classes in test data. The training data from previous NADI dialect identification tasks is multi-class. Hence, we must devise approaches to utilize the patterns or label information in these data to provide a multi-label classification for the test data. Further, the development set provided had only 8 dialects (multi-labeled), while in the test set, the number of dialects is unknown, with a maximum number of dialects specified as 18. This restricts us from using any label-specific statistical observational patterns derived from the development set to be adapted directly to the test set.

Under these challenging settings, we experimented with supervised and unsupervised approaches to derive multi-label predictions on the test set. For the final submission, we used the approaches that provided the best scores on the dev set.

3.1 Unsupervised Approaches

For the unsupervised settings, we considered the training data as a database and the input test sample as a query, hence modeling the task as a retrieval problem. The difference is that in this setup, the database is labeled; hence, the associated labels of the retrieved data samples can be considered related to the query. Considering the Training Set T_D and the input sample query Q, we retrieve *K* documents from T_D, forming the retrieved set R_D. Each document $d \in T_D$ is associated with a label class l. hence, the final retrieved set would be $(r_{d1}, l_1), (r_{d2}, l_2), .., (r_{dK}, l_K)$.

For retrieving the samples, we initially compute the sentence embeddings for all the samples in T_D. Sentence encoders (Reimers and Gurevych, 2019) using MARBERTV2 (Mageed et al., 2021), which was trained in Arabic and provided the best results in NADI-2023 shared tasks were used. As discussed, we treat each input sample from the dev/test set as a query. This sample query is also embedded. Further, we apply a semantic search (based on similarity) between each query and the encoded train set samples. Then, the top *K* similar train samples are retrieved to get the retrieved set R_D. The assumption is that the associated labels of these retrieved *K* samples can give us some notion of the labels associated with the query. Once we have the R_D set, the next step is to use the associated labels to label the query. We use different approaches to associate these labels with the given query. For all the approaches, we used *K=10*.

Unsupervised Cross Encoder based Label Count Threshold (Un-Cross-LCT) Approach: In this case, we use a heuristic-based approach based on the counts on the labels in R_D. First, we sort the labels in the descending order of the counts. If among the top 10, we have $l <= 3$ unique labels, we assign them to the query. If the most common label (based on count), i.e., Rank 1 in the sorted R_D, has a $count >= 5$, then assign only that label to the query. Otherwise, check for

the lower-ranked labels if $2 < count < 5$. Include them in the multi-label set if $1 < count <= 3$. For instance, consider that the label counts from *Prediction X* is : [('Palestine', 4), ('Algeria', 3), ('Yemen', 1), ('Sudan', 1). Applying the Un-Cross-LCT approach, we will associate the test sample with *Palestine* & *Algeria* as the labels.

We also extended this approach and experimented with variations using the post-processing methods explained in Section **??**.

3.2 Supervised Approaches

For the supervised approach, we used NADI-2023 data to fine-tune a pre-trained Arabic model. We used MARBERTV2[1] (Mageed et al., 2021) as our pre-trained model. MARBERT is a large-scale pre-trained masked language model focusing on Dialectal Arabic (DA) and MSA. It was trained on randomly sampled 1B Arabic tweets from a large in-house dataset of about 6B tweets. Further, a new version was released, trained on a bigger sequence length of 512 tokens for 40 epochs, i.e., MARBERTV2. The model predictions from the fine-tuned MARBERT model would be naturally multi-class. Thus, the next step would be enhancing the labels to a multi-label set-up described in Section 3.3. Another approach was to train a binary classifier for each dialect and finally get a label prediction from each classifier. The training was also based on NADI-2023 data and using MARBERTV2.

3.3 Post-Processing: Label Enhancement Approaches & Filtering

In this section, we describe the various post-processing steps. We used two main approaches: label enhancements and filtering. The former is recall-oriented, while the latter is precision-oriented. We also tried a combination of these approaches. The different post-processing steps are enlisted:

1. **Filtering (Post-FiltN)**: In this approach, we use filtering as the post-processing step. After obtaining the initial predictions from the supervised or unsupervised approaches, we tried out with different *N* values, where *N* represents how many multi-label predictions need to be kept. We computed the F-score on the dev set with different *N* values. For instance,

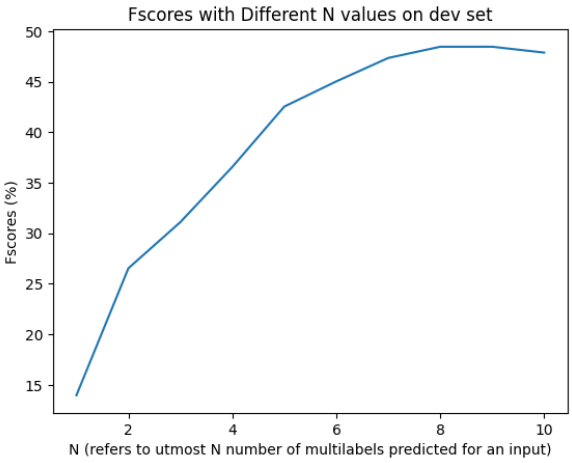

Figure 2: Fscores with Different N values for the Post-FiltN approach on dev set

Figure 2 plots *N* versus F-score (%) under an unsupervised setup. In this case, we get the top *K* predictions and then apply different N values. it can be observed that the maximum F-score is obtained at N=8 & 9, while at N=10, the F-score decreases.

Similar observations were made under the supervised setup.

2. **Co-occurrence Based (Post-Co)**: In this post-processing step, we tried to enhance the multi-labels using the label co-occurrences derived from the development set. We considered Pearson correlations between the labels and considered all co-occurrences with a correlation value, $r <= \theta$. In our experiments, we used a $\theta = 0.25$. For instance, the dialectal labels co-occurring with *Algeria* were *['Sudan', 'Tunisia']*. In this case, every time the approach predicts *Algeria*, by default, we also include *'Sudan'* and *'Tunisia'*.

3. **Confusion Matrix Based (Post-CM)**: In this case, we use the confusion matrix obtained using NADI-2023 dev data on NADI 2023 fine-tuned model. The idea is to use all the confused or misclassified labels for each dialect. The multi-class labels were then extended based on these CM-based label sets.

4 Experimental Settings and Results

For implementations, we used the models and libraries from HuggingFace and Sentence Trans-

[1] https://huggingface.co/UBC-NLP/MARBERTv2

Set-Up	Approach	Descriptions	Fscore (%)	Accuracy (%)
Unsupervised	Un-Cross-LCT	Cross-Encoder with Label Count Threshold	45.57	67.5
	Un-Cross+ Post-FiltN (N=8)	Cross-Encoder with Filtering	48.45	70.10
	Un-Cross + Post-Co+ FiltN (N=4)	Cross-Encoder with Co-occurring Label Enhancements & Filtering	48.45	70.10
Supervised	Sup (Mono)	Supervised with Multi-class	8.45	60.42
	Sup-Bin	Supervised Binary Classifier	19.21	42.61
	Sup-CM	Supervised Binary Classifier with Confusion Matrix based Label Enhancement	29.01	54.58
	Sup-Bin+Post-FiltN (N=8)	Supervised Binary Classifier with Filtering	19.21	42.61
	Sup-CM+Post-FiltN (N=9)	Supervised Binary Classifier with Confusion Matrix based Label Enhancement and Filtering	37.37	62.40
Baselines	Word_Jaccard	Word based overlaps with Jaccard	18.1	62.92
	BPE_Jaccard	BPE based overlaps with Jaccard	9.91	62.5
	BM25	BM25 IR approach	13.57	61.35

Table 1: Detailed Evaluation Results on Development Set

Approach	Fscore (%)	Accuracy (%)	Precision (%)	Recall (%)
Un-Cross-LCT	31.5	72.87	**64.0**	21.4
Un-Cross+ Post-FiltN (N=8)	**43.27**	71.88	53.64	**37.42**
Sup-CM+Post-FiltN (N=8)	29.41	59.01	32.15	33.13
Un-Cross+Post-Co+ FiltN (N=4)	31.81	65.56	42.0	32.75
Un-Cross+Post-Co+ FiltN (N=8)	43.08	58.42	39.29	57.21

Table 2: Evaluation results on test set

formers [2]. For the unsupervised approach, as described, we used the MARBERTV2 model with the Sentence Transformer library. For the cross-encoder part we used the multi-lingual version trained on mMARCO dataset (Bonifacio et al., 2021), which is the multilingual version of MS MARCO passage ranking dataset(Bajaj et al., 2016). For the supervised approach, to fine-tune the MARBERTV2 on NADI-2023 training data, we used a $\{dropout = 0.3, learningrate = 1e-5, batch\_size = 8, numberofepochs = 5\}$. The best model gave an F-score of 84% on the NADI-2023 dev set.

Table 1 reports the results of all the experiments conducted on the development set. As baselines, we also evaluated the development set using traditional n-gram-based and IR methods. We tried n-gram overlaps with Jaccard and Byte pair Encoding (BPE) based Jaccard overlaps. For the BPE experiments, we used the SentencePiece library [3] with a small vocab_size 500. The assumption was that initial merges could capture fine-grained linguistic nuances. Once the tokens are obtained, we remove the boundary tokens and consider all merges > 1 to find the overlaps. We also tried the simple IR algorithm, BM25 lexical retrieval model, M25 IR model, a ranking algorithm that determines document relevance to a query and ranks them based on a relevance score (Robertson et al., 1992), for comparisons. It can be observed that the best performing approach is with the unsupervised set-up with post-filtering, giving an F-score of 48.45%. Similar performance was obtained with a combination of co-occurrence based label enhancements. It can also be noted that the post-processing approaches facilitated performance improvement. We observed that our approaches provided better precision while recall was less. The best-performing approach presented a recall of 41.39%, with a precision of 63.72%.

Based on these results, we used the best models to evaluate the test data. The results are shown in Table 2. It can be observed that the best result is obtained with the unsupervised approach using post-filtering, presenting an F-score of 43.27% (Ranked 2), also giving the best recall. The best precision of 64% was obtained with the count-threshold based unsupervised approach. In general, it was noted that our approaches presented better precision, and recall was comparatively low. We submitted only one supervised approach with post-filtering and confusion matrix-based label enhancements, giv-

[2] https://github.com/UKPLab/sentence-transformers

[3] https://github.com/google/sentencepiece

ing an F-score of only 29.41%

A better recall could be obtained by adjusting the post-processing parameters such as filtering *N* values. However, this can decrease precision, and hence, we need to explore strategies to achieve a better recall-precision trade-off. As expected, the unsupervised approaches outperformed the supervised settings since the supervision was based on a dataset under multi-class settings. Considering the challenging setting, the proposed approach performed reasonably well.

5 Conclusion

In this paper, we described our experimental approaches for the NADI 2024 Subtask 1, which deals with multi-label dialect identifications. The absence of multi-label datasets poses a major challenge in this setup. The idea is to use multi-class datasets and adapt them for multi-label predictions. The complexity becomes multi-folded since the number of dialects in the test set was kept unknown. We use a cross-encoder-based approach in the unsupervised set-up with heuristics based on label count thresholds. Further, filtering and label enhancements were applied in the post-processing step. In the supervised setup, we used the trained classifiers for label predictions and then applied label enhancements or used binary classifiers for predictions. We observed that the cross-encoder-based approach with post-processing performed the best. A generalized conclusion of the best approach to MLDID cannot be made without sufficient properly annotated data. Hence, continued efforts in this direction must be made while experimenting with suitable multi-class adaptation strategies.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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, 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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in arabic broadcast speech.

Mohamed Ali. 2018. Character level convolutional neural network for arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 122–127.

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting marbert for improved arabic dialect identification: Submission to the nadi 2021 shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 260–264.

Nouf AlShenaifi and Aqil Azmi. 2020. Faheem at nadi shared task: Identifying the dialect of arabic tweet. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 282–287.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Andrei Butnaru and Radu Tudor Ionescu. 2018. Unibuckernel reloaded: First place in arabic dialect identification for the second year in a row. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 77–87.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-oslo team at the vardial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65.

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.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Bert-based multi-task model for country and province level msa and dialectal arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.

Salima Harrat, Karima Meftouh, Karima Abidi, and Kamel Smaïli. 2019. Automatic identification methods on a corpus of twenty five fine-grained arabic dialects. In *International Conference on Arabic Language Processing*, pages 79–92. Springer.

Muhammad Abdul Mageed, Abdelrahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the third workshop on NLP for similar languages, varieties and dialects (VarDial3)*, pages 1–14.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen E Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. 1992. Okapi at trec. In *Text retrieval conference*, pages 21–30.

Samia Touileb. 2020. Ltg-st at nadi shared task 1: Arabic dialect identification using a stacking classifier. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 313–319.

Mutaz Younes, Nour Al-Khdour, and AL-Smadi Mohammad. 2020. Team alexa at nadi shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 237–242.

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.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging. the second vardial evaluation campaign.