ELYADATA at NADI 2024 shared task: Arabic Dialect Identification with Similarity-Induced Mono-to-Multi Label Transformation

Amira Karoui[1] , **Farah Gharbi**[1], **Rami Kammoun**[2], **Imen Laouirine**[1] and **Fethi Bougares**[2]
ELYADATA[1] , ALGOBRAIN[2]
firstname.lastname@elyadata.com[1], firstname.lastname@algobrain.ai[2]

Abstract

This paper describes our submissions to the Multi-label Country-level Dialect Identification subtask of the NADI2024 shared task, organized during the second edition of the ArabicNLP conference. Our submission is based on the ensemble of fine-tuned BERT-based models, after implementing the Similarity-Induced Mono-to-Multi Label Transformation (SIMMT) on the input data. Our submission ranked first with a Macro-Average F1 score of 50.57%.

Keywords: Ensemble method, Multi-Label, BERT-based models.

1 Introduction

Dialect Identification (DI) is the task of automatically determining the specific dialect (country) or regional variation that a text belongs to. In recent years, NADI[1] has been dedicated to provide diverse datasets and modeling opportunities to advance Arabic Natural Language Processing (NLP), including Arabic dialects. NADI subtasks are not limited to DI, it includes other NLP tasks such as sentiment analysis (Abdul-Mageed et al., 2022), etc... There are 3 subtasks for NADI's competition this year, we chose to participate in Multi-label country-level Dialect Identification (MLDID). This shared task started from 2020, and the DI task has been consistent in all previous editions of NADI. It has shown a significant improvement over the past years. As an instance, in 2020, the best system for the DI task, achieved a macro-average F1 score of 26.78% (Talafha et al., 2020), while in 2023, the macro-average F1 score increased to 87.27% (El-karef et al., 2023).

However, this year represents an exception for the shared subtask, NADI extended the challenge to cover the task of MLDID which is identifying and categorizing regional or social variations in a language. When looking at today's research literature, as far as we know, there are no published systems on MLDID for Arabic dialects. When trying to solve such a task, multiple challenges rise: (1) It is very difficult for native speakers of one country to identify other dialects and whether a sentence belongs to multiple ones or not (Malmasi et al., 2015). (2) There are not a fully conventional standard writing in each dialect used by its native speakers which adds to the complexity (Abdallah et al., 2023), especially when the texts are scraped from the internet where each individual writes with their own style. (3) The code switching existing in each dialect, especially in French or English depending on the region, where each person writes foreign words in Arabic (Hijjawi and Elsheikh, 2015). (4) Some cross-dialectal words are written in the same way which makes it hard to identify to which dialect it belongs (Tachicart et al., 2022).

The main contributions of this work can be summarized as follows:

- Converted the training data from mono to multi-labeled by applying the method of vocabulary similarity.
- Utilized Ensemble methods on fine-tuned state of the art text classification models.
- Achieved first place for the MLDID subtask in NADI2024.
- Made all Pre-processing and modeling scripts available at [2].

This paper is organized as follows: Section 2 presents the dataset and its peculiarities. In section 3, approaches towards building our systems were detailed. Section 4 discusses the experiments

[1] https://nadi.dlnlp.ai/

[2] https://github.com/elyadata/NADI_shared_task_2024

Proceedings of The Second Arabic Natural Language Processing Conference, pages 758–763
August 16, 2024 ©2024 Association for Computational Linguistics

conducted in order to build the systems and the achieved results. Finally, we finish our paper by a discussion and a conclusion.

2 Data

The provided NADI dataset is balanced and mono-labled accross 18 Arabic dialects. Each class consists of 1000 tweets where the average tweet length comprises 16 words. It is important to note that it is the same training set as the NADI-2023 dataset. Table 1 reports the dataset distribution over train, dev and test sets.

Splits	Sentences	Classes
Train	18000	18
Dev	120	8
Test	1000	Unknown

Table 1: Summary of Dataset Splits

NADI-2020 and NADI-2021 datasets were also provided by the shared task organizers. These datasets exhibit distinctions from the NADI-2023 dataset, characterized by uneven label distribution. Notably, Bahraini and Qatari dialects were underrepresented compared to the predominantly represented Saudi Arabian and Egyptian dialects. However, it should be noted that the tweets were classified based solely on their posting locations, which may not always be precise.

The new addition in this year's competition is a dev set introduced for the multi-label classification (Maclin and Opitz, 2011) task. This dev set comprises of 120 entries spread over 8 classes: Algeria, Egypt, Jordan, Palestine, Sudan, Syria, Tunisia, and Yemen.

3 Systems

In this work, several approaches were experimented. Some of them are data-centric, such as data preprocessing and SIMMT, while others are model-centric, including Ensemble methods, Binary Relevance, and Staged Fine-tuning.

But before delving into these approaches, the first step was to choose the model architecture to work with. BERT-based models (Devlin et al., 2018), which are Masked Language Models (MLM) (Taylor, 1953), seemed the most adequate. Since the dataset is in DA, language models that have already been trained on Arabic were explored.

3.1 Dataset preprocessing

The NADI-2023 dataset comprised raw tweets, containing noise such as Arabic laughter expressions like 'XD', 'ههههه' (hhhh) and similar variations along with diacritics like 'نگتب' (We type). To address this issue, cleaning and normalization procedures were conducted (Lichouri et al., 2023). Hashtags, single letters, diacritics, and laughter expressions were removed. This resulted in a standardized dataset prepared for model training.

3.2 Ensemble Methods

Experimentation with Ensemble methods (Yang et al., 2023) was also conducted to further enhance our system's performance. Two approaches were tested, the first was the Average Classifier (Mohammed and Kora, 2023), it combines predictions by averaging the probabilities for each class from all models in the ensemble. The second approach, which is the maximum probability ensemble (Kundu et al., 2021), compares probabilities from three models for events. Each model assigns probabilities independently. Then, the maximum probability for each event across all models is determined, combining predictions to emphasize events with the highest collective confidence.

3.3 Similarity-Induced Mono-to-Multi Label Transformation

To enhance the model performance, given the multilabeled nature of the dev and test sets, one of our key approaches involved transforming the dataset from mono-label to multi-label using a SIMMT technique.

To implement this transformation, a vocabulary of dialect-specific words from the dataset was created. Then, the exact similarity score of words between each sentence and the vocabulary of each of the 18 dialects was calculated. The transformation process followed a binary assignment:

$$C_{ij} = \begin{cases} 1, & \text{if } \mathrm{sim}(S_j, D_i) > \textit{threshold} \\ 0, & \text{otherwise} \end{cases}$$

Using a threshold-based approach, sentences were evaluated against the vocabulary of each dialect. Sentences that exceeded a specific similarity threshold were assigned to the corresponding class, while those that did not were excluded.

3.4 Other Techniques

3.4.1 Binary Relevance

As the task is for multi-label classification, the idea is to train 18 classifiers (Aldrees et al., 2016), each on a dialect, then combine their results. Two approaches were adapted: balanced and unbalanced training. The first includes training 1,000 sentences as a 'yes' label and the 17,000 rest as a 'no' label. The second tries to reach a balance in training data by randomly picking 1,000 samples from the other dialects and the full 1,000 from the targeted dialect when training each classifier.

Abdul-Mageed et al. (2020) has proved that in dialectal assessments, using MarBert is more efficient than ArBERT. TunBert(Messaoudi et al., 2021), DarijaBert(Gaanoun et al., 2024) and DziriBert (Abdaoui et al., 2021) were fine-tuned for Tunisian, Moroccan, and Algerian, respectively. For the rest, a MarBERT was fine-tuned on the targeted dialect.

3.4.2 Staged Fine-Tuning

As already mentioned in the section 2, the training sets for the previous NADI editions were provided. Here comes the staged fine-tuning idea introduced in El-karef et al. (2023), which involves fine-tuning the model on data from previous years and then fine-tuning it on the current data. To execute this idea, our model was fine-tuned three times on three distinct datasets: NADI-2020 for the first fine-tuning, NADI-2021 for the second fine-tuning, and NADI-2023 for the last, with each, the same methods of preprocessing were applied.

4 Experiments and Results

As already mentioned in section 3, BERT-based models trained on Arabic language were chosen as base models for our experiments. Several pre-trained models were tested, mainly: CAMeLBERT (Obeid et al., 2020), MarBERT, ArBERT, MarBERTV2 (Abdul-Mageed et al., 2020), and their variations.

To ensure equality amongst experimented systems, the same hyper-parameters in Table 2 were kept for all experiments. After several experiments, the following models gave the best results: MarBERT, ArBERT and MarBERTV2. Moving forward, the latter models were chosen.

4.1 Dataset Preprocessing

To evaluate the impact of the preprocessing methods on model performance, the chosen models were fine-tuned using both the original and preprocessed versions of the NADI-2023 mono-labeled dataset.

Hyper-parameter	Value
Learning Rate	1e-05
Optimizer	AdamW
Train Batch Size	11
Evaluation Batch Size	11
Number of Training Epochs	10
Dropout Rate	0.3

Table 2: Fine-Tuning Hyper-parameters.

Table 3 illustrates the comparative performance of various models on both preprocessed and non-processed datasets. Notably, the experiment achieved the best results with the MarBERTv2 model, particularly on the processed dataset.

Model	Preprocessed	Not Processed
MarBERT	0.080	0.076
ArBERT	0.090	0.084
MarBERTv2	0.091	0.086

Table 3: The effect of pre-processing on the listed fine-tuned models results

This low results shown could be explained by the fact that the sentences in the dev set are multi-labeled while our model predicts only one label for each sentence.

4.2 Ensemble

Regarding the ensemble approach, ArBERT, MarBERT and MarBERTv2 were fine-tuned with the same settings listed in table 2.

We experimented with 2 ensemble methods: The average classifier and the max probabilities. In the former the classifer averages the outputted probabilities of each model then applies a Ensemble Threshold (ET) for each class to determine the labels. Whereas in the latter, each model generates a probability for each class, only the ones higher than a certain threshold were needed to be chosen. As shown in table 4 the average classifier achieved a macro-average F1 score of 0.574, while it was slightly outperformed by the Max Probabilities Ensemble (MPE) with a score of 0.580.

Based on the observed improvement of the model performance with the MPE, it was decided to move forward with it. Various ensemble thresholds were experimented to enhance the model's

Approach	ET	Macro-Avg F1
Average classifier	0.0003	0.574
Max probabilities	0.0004	0.580

Table 4: Comparison between Average classifier and Max probabilities approaches using different Ensemble thresholds.

predictions. This phase revealed a macro-average F1 Score of 0.347, 0.340, and 0.3354 for the thresholds of 0.3, 0.4, and 0.5, respectively. However, as the threshold values of the probabilities were gradually decreased, reaching as low as 0.0002, a significant improvement in performance became apparent, indicating that 0.0002 was the optimal threshold value, improving the performance by 0.27 macro-average F1 score.

This low threshold was chosen since upon observing the output probabilities for each sentence, big difference was noticed. One label had a strong probability (more than 0.9), but the rest label probabilities had very low values. This threshold adjustment addresses the big gap between the probabilities.

4.3 Mono-labeled to multi-labeled dataset transformation

After extracting the vocabulary from the dataset, the next step was to determine the optimal similarity threshold to evaluate the similarity between each sentence and the vocabulary of each dialect. By setting the similarity threshold to 60% a macro-average F1 score of 0.588 was achieved, then, when decreased the similarity threshold to 40% a macro-average F1 score of 0.6 was obtained. 30% was also tested but the results were lower. So, 40% was chosen as a similarity threshold.

Additionally, our approach was further refined through code implementation. In cases where one or two labels were zero while all other labels were one, a value of 1 was assigned to all labels. This adjustment successfully identified features belonging to MSA and improved our model's performance, raising the macro-average F1 score from 0.608 to 0.6108.

4.4 Other Techniques

4.4.1 Binary Relevance

The initial unbalanced experiment gave a macro-average F1 Score of 0.49 which compelled us to move to the balanced approach. For the latter, the instances were randomly shuffled to ensure randomness. However, despite these efforts, the macro-average F1 score dropped to 0.17, which could be attributed to the low amount of data for each dialect.

It was also observed that the Tunisian classifier gave the worst labeling since its initial training data composed mostly of Latin alphabets.

Moreover, what added to the general failure of this experiment to attain proper results can be contributed to the fact that if each classifier had mistaken its prediction, consequently, the overall juncture of the predictions would follow on that path of wrong labeling.

4.4.2 Staged Fine-Tuning

Regarding the staged fine-tuning, as already mentioned in section 3.4.2, there are three steps, and for all stages, all the hyper-parameters are the same as those presented in Table 2. Concerning the input data, the same preprocessing described in section 3.1 was applied for all stages. This experiment yielded a macro-average F1-Score of 55% on the dev2 set, so it did not surpass the achieved performance. Therefore, this system was aborted.

4.5 Submitted systems

The system which yielded the best results on the dev2 set described in section 4.3 was our primary submission. It achieved the highest rank among all participants with a macro-average F1 score of 50.57% on the test set.

5 Discussion

After analyzing the results and exploring the dev set further, problems with wrongly labeled phrases were found. Some labels were mistakenly assigned to many instances, creating inconsistencies in the dataset. For example, one sentence labeled as palestine "لو سمحت ابي واحد شاي" (Please I want one cup of tea), reflects a linguistic nuance, using "ابي" (I want), which is more a characteristic of Gulf Arabic rather than Palestinian Arabic.

Another challenge faced was determining labels for transliterated sentences, like "اقولها لتس فيس تو فيس افتر ديز" (I say let's face to face after this), which can be confusing even for native speakers to classify.

Mistakes in labeling do not just add noise but also make it hard to judge how well a model performs accurately. It is crucial to refine data annotation procedures and implement rigorous quality checks to mitigate the impact of mislabeled data on model training and evaluation.

6 Conclusion

This paper presents our team's submission to the Multi-label country-level Dialect Identification of the 2024 NADI shared task. Our submission relies on the usage of Similarity-Induced Mono-to-Multi Label Transformations a data-centric approach and Max Probabilities Ensemble on fine-tuned models which is a model-centric approach. This allowed us to be ranked first in the official evaluation with a macro-average f1 score of 50.57%.

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2023. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. *Preprint*, arXiv:2309.11327.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *CoRR*, abs/2109.12346.

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

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

Asma Aldrees, Azeddine Chikh, and Jawad Berri. 2016. Comparative evaluation of four multi-label classification algorithms in classifying learning objects. *Computer Science Information Technology*, 6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mohab El-karef, Movina Moses, Shinnosuke Tanaka, James Barry, and Geeth Mel. 2023. Nlpeople at nadi 2023 shared task: Arabic dialect identification with augmented context and multi-stage tuning. In *Proceedings of ArabicNLP 2023*, pages 642–646.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, pages 1–13.

Mohammad Hijjawi and Yousef Elsheikh. 2015. Arabic language challenges in text based conversational agents compared to the english language. *International Journal of Computer Science and Information Technology*, 7:1–13.

Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. 2021. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. *PloS one*, 16(9):e0256630.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. Usthb at nadi 2023 shared task: Exploring preprocessing and feature engineering strategies for arabic dialect identification. *arXiv preprint arXiv:2312.10536*.

Richard Maclin and David W. Opitz. 2011. Popular ensemble methods: An empirical study. *CoRR*, abs/1106.0257.

Shervin Malmasi, Eshrag A. Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *International Conference of the Pacific Association for Computaitonal Linguistics*.

Abir Messaoudi, Ahmed Cheikhrouhou, Hatem Haddad, Nourchene Ferchichi, Moez BenHajhmida, Abir Korched, Malek Naski, Faten Ghriss, and Amine Kerkeni. 2021. Tunbert: Pretrained contextualized text representation for tunisian dialect. *CoRR*, abs/2111.13138.

Ammar Mohammed and Rania Kora. 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Ridouane Tachicart, Karim Bouzoubaa, Salima Harrat, and Kamel Smaïli. 2022. *Arabic Dialects Morphological Analyzers: A Survey*, pages 189–203. Springer International Publishing, Cham.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Yongquan Yang, Haijun Lv, and Ning Chen. 2023. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, 56(6):5545–5589.