Frank at NADI 2023 Shared Task: Trio-Based Ensemble Approach for Arabic Dialect Identification

Dilshod Azizov[1], Jiyong Li[2], Shangsong Liang[1,*]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Sun Yat-sen University
dilshod.azizov@mbzuai.ac.ae, lijy373@mail2.sysu.edu.cn
[*]Corresponding author, liangshangsong@gmail.com

Abstract

We present our system designed for Subtask 1 in the shared task NADI on Arabic Dialect Identification, which is part of ARABICNLP 2023. In our approach, we utilized models such as: MARBERT, MARBERTv2 (A) and MARBERTv2 (B). Subsequently, we created a majority-voting ensemble of these models. We used MARBERTv2 with different hyperparameters, which significantly improved the overall performance of the ensemble model. In terms of performance, our system achieved a competitive an F1 score of **84.76**. Overall, our system secured the 5th position out of 16 participating teams.

1 Introduction

The Arabic language, with its vast and varied tapestry of dialects, offers a mesmerizing blend of history, culture and linguistic evolution. Each dialect, from the mellifluous notes of Levantine to the rhythmic cadences of Maghrebi, narrates a unique story of its people, their journeys, and their experiences. However, such linguistic richness often goes unnoticed, overshadowed by mainstream dialects and a lack of comprehensive research tools. The persistent gaps in our understanding, exacerbated by limited resources, such as datasets, have made the exploration of these dialects both a challenge and a treasure hunt for researchers (Althobaiti, 2020).

In response to this, the series of nuanced Arabic dialect identification (NADI) shared tasks, initiated by (Abdul-Mageed et al., 2020b), emerged as a beacon of hope, spotlighting lesser studied dialects. Over the years 2020 (Abdul-Mageed et al., 2020b), 2021 (Abdul-Mageed et al., 2021), and 2022 (Abdul-Mageed et al., 2022), NADI provided invaluable datasets and created a vibrant platform where scholars and enthusiasts could exchange insights, challenge conventional methodologies, and ignite renewed interest in dialect identification. This discipline, which is based on determining the variety of sources of textual or spoken content, has now become central to understanding the rich fabric of the Arabic linguistic diversity.

The subtask can be formulated as follows:

Identify the specific country-level dialect of a given Arabic tweet.

This task is armed with the novel TWT-2023 dataset, which covers 18 mesmerizing dialects, and is supplemented by external datasets such as NADI-2020-TWT, NADI-2021-TWT and MADAR-2018 (Bouamor et al., 2018).

Our contributions are as follows:

- We propose an automated system based on the majority-voting ensemble that uses MARBERT (Abdul-Mageed et al., 2020a), MARBERTv2 (A) and MARBERTv2 (B) for the Dialect Identification.
- We compare the performance of MARBERT, MARBERTv2 (A) and MARBERTv2 (B).

In Section 2, we outline previous and more recent studies on dialect identification. In Section 3, we illustrate a thorough examination of the dataset. In Section 4 we describe the system and the results. Lastly, Section 5 presents our conclusion and proposes potential avenues for future research.

2 Related Work

Arabic exists in three main forms: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). Although CA and MSA have been thoroughly explored in previous research, interest in DA has recently risen due to limited resources (Holes, 2004; Brustad, 2000).

The initial research on DA was regional (Gadalla and ElMaraghy, 1997; Diab et al., 2010), later expanding to multi-dialectal studies (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014;

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 637–641

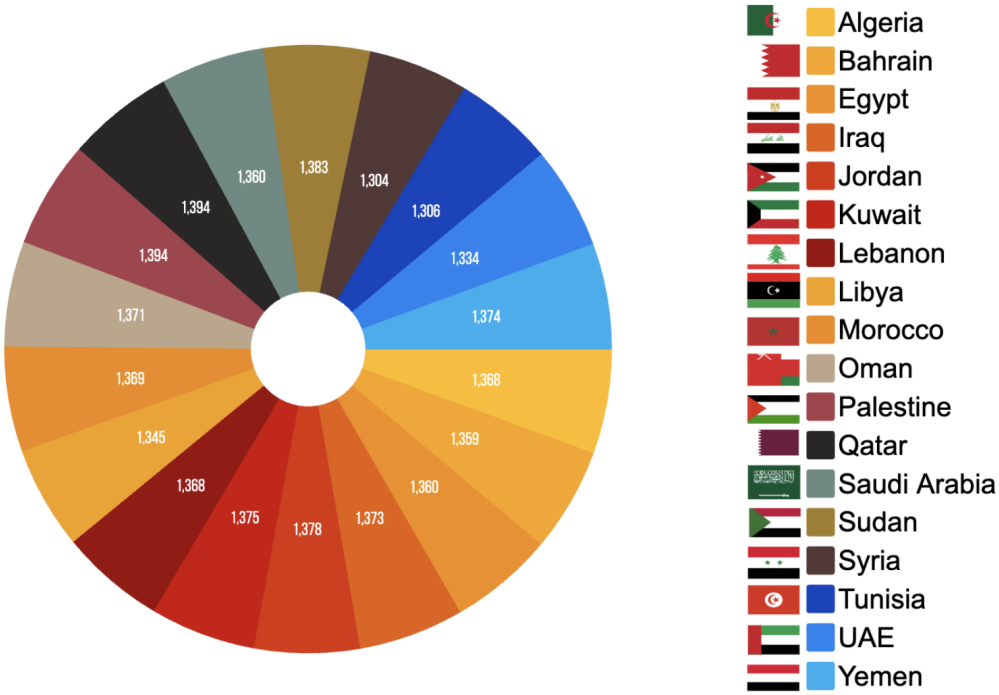

Figure 1: Statistics about tweets distribution in train and development sets.

Bouamor et al., 2014). The VarDial workshop highlighted the identification of dialects using acoustic and phonetic traits (Zampieri et al., 2017).

MADAR (Bouamor et al., 2018) provided enriched dialect data, but faced authenticity questions in online contexts. Recent work has taken advantage of the vast Twitter datasets (Mubarak and Darwish, 2014; Abdelali et al., 2021), with Althobaiti (2022) introducing an unsupervised dialect-tagging approach. Further, Abdul-Mageed et al. (2020b) investigated city-specific dialect variations.

NADI's initiatives produced notable datasets on Arabic dialect identification, including a detailed review by Althobaiti (2020). NADI 2020 collaborated with WANLP 2020, leading to the categorization of dialects from 21 Arab countries via Twitter. NADI 2021, in association with WANLP 2021, improved its dataset, distinguishing between MSA and DA. This led to the development of four specific subtasks (Abdul-Mageed et al., 2021). In NADI 2022, the focus had shifted to sentiment analysis of data tagged with dialects. In particular, Alsudais et al. (2022) integrated the MADAR and NADI datasets into their research. Lastly, NADI 2023 introduced three subtasks: country-level dialect identification and closed- and open-speech machine translation from four dialects to MSA.

3 Data

This section provides a detailed explanation of the dataset made available by the NADI shared task organizers.

Data Attributes:

- **ID:** A numerical index assigned to each data point.
- **Tweet:** An Arabic tweet written in various dialects.
- **Label:** Indicates the specific dialect corresponding to one of the 18 countries (e.g., UAE, Morocco, etc.).

Dataset Size:

The statistics of the dataset for this task are detailed in Figure 2. In total we have slightly more than 28K. We used an external dataset from the set, which is provided by organizers (NADI-2021-TWT). The distribution of labels within the training and development sets can be seen in Figure 1. In particular, the dataset has a balanced distribution.

4 System Description and Results

4.1 System Description

For evaluation, we use the official evaluation scorers provided for the shared task. The primary measure for our subtask is an F1 score. Our model was executed on 2 NVIDIA Tesla T4 (16GB) GPU.

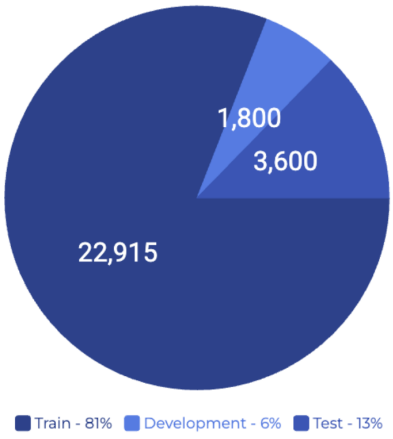

Figure 2: Statistics about tweets distribution in train/development/test sets.

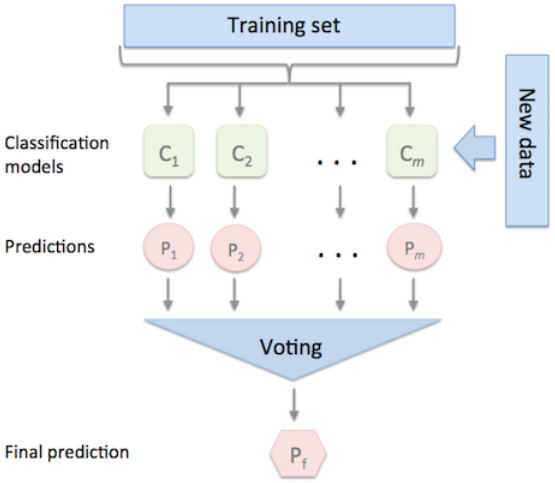

Figure 3: Majority voting architecture. Source: www.researchgate.net

We take advantage of the majority vote technique in ensemble learning as an alternative method (Dietterich et al., 2002; Sagi and Rokach, 2018; Zhu et al., 2021). We opted for the majority-voting ensemble due to our balanced dataset. This technique aggregates predictions from multiple models for a given input. The architecture is shown in Figure 3, where the final prediction is derived from the class or result that receives the majority vote from the ensemble (Da San Martino et al., 2023; Azizov et al., 2023; Barrón-Cedeño et al., 2023a,b).

Consider m classifiers, $C_1, C_2, \ldots, C_m$, predicting the class label for an input x as $P_1, P_2, \ldots, P_m$. The majority-voting classifier gives the final class label, P_f, based on the most frequent prediction:

$$P_f = \text{mode}(P_1, P_2, \ldots, P_m) \quad (1)$$

For our task, we opt for hard voting, addressing concerns of classifier calibration and avoiding potential overconfidence in predictions. This ensures that the majority consensus dictates the final prediction. Although our method relies on the most reliable framework in the case of varying model predictions.

The following is the experimental setup for our models:

MARBERT: This model was trained for 1 epoch using a learning rate of 5e-5 and a weight decay of 0.001.

MARBERTv2 (A): MARBERTv2 was trained for 2 epochs with a weight decay of 0.0.

MARBERTv2 (B): This version of MARBERTv2 was trained for 2 epochs with a weight decay of 0.001.

Unless specified otherwise, all other hyperparameters were kept at their default values.

All these mentioned models were combined using the architecture shown in Figure 3. In case of differing predictions across all three models, we prioritize the prediction MARBERTv2 (B) due to its superior performance.

To maximize performance, we used a customized training approach in our study with three model versions (MARBERT, MARBERT A, and MARBERT B). The models showed inherent similarities, but different optimal training epochs were identified: MARBERT peaked at the first epoch, whereas both MARBERT A and MARBERT B performed optimally in the second epoch. To avoid overfitting, training was stopped in these instances.

4.2 Results

In this section, we discuss the results of our models.

We experimented with the development set, since we used it as a test set, and from the train set we cut 10% out of the total tweets for the development set.

MARBERT vs. MARBERTv2 (A): A comparison between the original MARBERT model and its first variant MARBERTv2 (A) shows noticeable improvements in all measures in the latter. An F1 score sees an increase of 1.99 percentage points, moving from 82.40 in MARBERT to 84.39 in MARBERTv2 (A). Similarly, the precision in MARBERTv2 (A) is higher by 2.14 percentage points than the original MARBERT, which is 84.73.

MARBERTv2 (A) vs. MARBERTv2 (B):

	F1 Score	Accuracy	Precision	Recall
MARBERT	82.40	82.38	82.59	82.38
MARBERTv2 (A)	84.39	84.33	84.73	84.33
MARBERTv2 (B)	84.44	84.38	84.72	84.38
Majority voting	**85.90**	**85.83**	**86.12**	**85.83**

Table 1: Experimental results of our frameworks on development set.

When comparing the two versions of MARBERTv2, the improvements in the (B) version, although modest, are discernible. An F1 score is marginally better by 0.05 percentage points in the (B) version. The precision in MARBERTv2 (B) is nearly the same as its counterpart (A), but sees a tiny decrease of 0.01 percentage points. This suggests that the adjustments made between the two versions of MARBERTv2 led to slight improvements in certain areas, but had a negligible impact on precision.

MARBERTv2 (B) vs. Majority Voting: The ensemble model, using a majority voting approach, clearly outshines the best performing MARBERTv2 version. An F1 score in the majority voting approach is higher by a significant 1.46 percentage points compared to MARBERTv2 (B). The precision is also improved in the majority voting method by 1.4 percentage points, making it the most precise model among the ones evaluated.

Overall Observations: Across the board, each subsequent version of the model or approach appears to bring about performance improvements, with the majority-voting method standing out as the most effective.

Based on the leaderboard results, we secured the fifth rank. Our achieved an F1 score is 84.76. For other evaluation measures, we recorded an accuracy of 84.75, a precision of 84.95, and a recall of 84.75.

5 Conclusion and Future Work

In this paper, we discussed our approach for subtask 1 of the shared task NADI in Arabic Dialect Identification. We used the majority-voting ensemble with the MARBERT and MARBERTv2 (A) and MARBERTv2 (B) models and according to the official leaderboard results, our system achieved an F1 score of **84.76** outperforming two-thirds of the participating teams. We also detailed a series of experiments and made comparisons of our models with a majority-voting ensemble.

In future work, we plan to enhance our ensemble approach with advanced transformer architectures (e.g., mBERT and XLM-RoBERTa) and data augmentation specific to Arabic dialects (e.g., back-translation or dialectical synonym replacement). Moreover, we would like to investigate classifier calibration and soft voting.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

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

Abdulkareem Alsudais, Wafa Alotaibi, and Faye Alomary. 2022. Similarities between arabic dialects: Investigating geographical proximity. *Information Processing & Management*, 59(1):102770.

Maha J Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*.

Maha J Althobaiti. 2022. Creation of annotated country-level dialectal arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Dilshod Azizov, S Liang, and P Nakov. 2023. Frank at checkthat! 2023: Detecting the political bias of news articles and news media. *Working Notes of CLEF*.

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, et al. 2023a. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *European Conference on Information Retrieval*, pages 506–517. Springer.

Alberto Barrón-Cedeño, Firoj Alam, Andrea Galassi, Giovanni Da San Martino, Preslav Nakov, Tamer Elsayed, Dilshod Azizov, Tommaso Caselli, Gullal S Cheema, Fatima Haouari, et al. 2023b. Overview of the clef–2023 checkthat! lab on checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 251–275. Springer.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kristen Brustad. 2000. *The syntax of spoken Arabic: A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Georgetown University Press.

Giovanni Da San Martino, Firoj Alam, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Preslav Nakov. 2023. Overview of the clef-2023 checkthat! lab task 3 on political bias of news articles and news media. *Working Notes of CLEF*.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74. Citeseer.

Thomas G Dietterich et al. 2002. Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. Aida: Identifying code switching in informal arabic text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 94–101.

Mohamed AE Gadalla and Waguih H ElMaraghy. 1997. Improving the accuracy of machined parametric surfaces using cutting force synthesis and surface offset techniques. In *ASME International Mechanical Engineering Congress and Exposition*, volume 26782, pages 181–187. American Society of Mechanical Engineers.

Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

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.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

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.

Matteo Zampieri, Andrej Ceglar, Frank Dentener, and Andrea Toreti. 2017. Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environmental Research Letters*, 12(6):064008.

Yadong Zhu, Xiliang Wang, Qing Li, Tianjun Yao, and Shangsong Liang. 2021. Botspot++: A hierarchical deep ensemble model for bots install fraud detection in mobile advertising. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–28.