ISL-AAST at NADI 2023 shared task: Enhancing Arabic Dialect Identification in the Era of Globalization and Technological Progress

Shorouk Adel
Arab Academy for Science,
Technology, and Maritime Transport,
Alexandria, Egypt
shoroukadel@student.aast.edu

Noureldin Elmadany
Arab Academy for Science,
Technology, and Maritime Transport,
Alexandria, Egypt
nourelmadany@aast.edu

Abstract

Arabic dialects have extensive global usage owing to their significance and the vast number of Arabic speakers. However, technological progress and globalization are leading to significant transformations within Arabic dialects. They are acquiring new characteristics involving novel vocabulary and integrating linguistic elements from diverse dialects. Consequently, sentiment analysis of these dialects is becoming more challenging. This study categorizes dialects among 18 countries, as introduced by the Nuanced Arabic Dialect Identification (NADI) shared task competition. The study approach incorporates the utilization of the MARBERT and MARBERT v2 models with a fine tunning processes. The findings reveal that the most effective model is achieved by applying averaging and concatenation to the hidden layers of MARBERT v2, followed by feeding the resulting output into convolutional layers.Furthermore, employing the ensemble method on various methods enhances the model's performance. Our system secures the 6th position among the top performers in the First subtask, achieving an F1 score of 83.73%.

1 Introduction

The Arabic region encompasses numerous cultures, each characterized by dialectal variations influenced by historical, geographical, and sociopolitical factors (Bouamor et al., 2014). While this variety showcases the region's cultural wealth, it creates difficulties when analyzing Arabic information, especially on social media networks. Moreover, the rapid evolution of the language in the digital age and the widespread use of social media are presenting a new era for the Arabic language. Modern communication tools are enabling speakers of various Arabic dialects to interact globally. This interaction is leading to a dynamic evolution of the language, characterized by the emergence of new vocabulary, slang, and expressions (Darvin, 2016). The continuous generation of new words and language adaptations is presenting a unique challenge for linguistic analysis. Therefore, Modern Standard Arabic has a disequilibrium between preserving tradition and adjusting to the demands of modern communication. Moreover, Arabic dialect identification plays a pivotal role in understanding regional language variations on social media. Improving this task has implications for cultural preservation, social analysis, and natural language processing technology. However, the presence of diverse Arabic dialects with distinct linguistic traits can pose challenges in analyzing and interpreting social media content (Salameh et al., 2018). People from different regions might use completely different words to express the same concepts.

Recent advancements in Arabic Dialect Identification research have been notable, with various studies addressing the intricate nuances of Arabic dialects. The MADAR shared task on fine-grained dialect identification (Bouamor et al., 2019) delved into sub-dialect distinctions, highlighting the complexity of Arabic language variations. Machine Translation of Arabic Dialects (Salloum, 2018) focused on adapting translation models to handle dialect-specific expressions, facilitating communication across dialect differences. Moreover, efforts in the Automatic Identification of Arabic Dialects in Social Media (Sadat et al., 2014) utilized natural language processing and machine learning to automate dialect recognition, revealing regional language trends online. Various methods, including feature extraction and machine learning algorithms (Zaidan and Callison-Burch, 2014), have contributed to improving automated dialect identification accuracy and uncovering the rich diversity of Arabic dialects. In the recent NADI shared task series (Abdul-Mageed et al.,

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 631–636

2020b, 2021, 2022), teams employed a range of approaches, including traditional methods like SVM with TF-IDF (Nayel et al., 2021), customized Bert-based models (AlKhamissi et al., 2021), and deep learning techniques with models like MARBERT and AraBERT (Messaoudi et al., 2022; Abdel-Salam, 2022; Attieh and Hassan, 2022). These efforts collectively contribute to the advancement of Arabic dialect identification, showcasing diverse methodologies and approaches in the field.

In this research, we aim to enhance the F1 score of Arabic dialect identification, provided by NADI shared task 2023 (Abdul-Mageed et al., 2023), by investigating the impact of various model enhancements. our study conducts a series of experiments using MARBERT and MARBERT v2 models (Abdul-Mageed et al., 2020a), involving various techniques. This approach includes concatenating hidden layers (Devlin et al., 2018) and processing the resulting outputs using CNN layers (Jacovi et al., 2018), BILSTM models (Graves et al., 2005), or a combination of BILSTM and CNN. Additionally, Experiments involve adapters with the MARBERT model (Pfeiffer et al., 2020). Finally, to maximize our results, our work utilizes ensemble methods that combine the outcomes of the majority of these experiments (Re and Valentini, 2012).

The rest of the paper is organized as follows: providing the dataset and its preparation are presented in Section 2. In Section 3, we explain the methodologies employed for Arabic Dialect Identification. Subsequently, Section 4 presents the results of our model's performance, including an analysis of our findings. In Section 6, we summarize and conclude our findings.

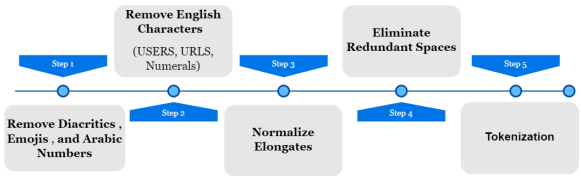

Figure 1: Pre-processing steps on the dataset

2 Data

2.1 Dataset Description

The presented approach utilized the training and validation data provided by the task organizers (Abdul-Mageed et al., 2023). The training set for Subtask 1 consists of around 18,000 tweets with 18 different labels representing 18 country dialects. While the development set consists of 1,800 labeled tweets. The submitted results were evaluated on a test set consisting of 3,600 tweets covering 18 country-level dialects.

2.2 Dataset Pre-processing

The dataset pre-processing is shown in Figure 1. The initial steps involved removing diacritics, which are modifications to Arabic characters. Subsequently, specific words were removed, such as mentions of users, URLs, and numerical values. Additionally, elongated characters were normalized to a single occurrence of the represented character. Emojis were also removed from the text. To further enhance the text, a series of processes were applied. Non-Arabic characters and redundant spaces were eliminated. Stemming or Lemmatization was not performed due to the intricacies of the Arabic language. These linguistic intricacies include the rich morphology and variability in Arabic dialects, where words may undergo significant changes in form and structure. Applying stemming or lemmatization involves reducing words to their root or base form. By observation, it could potentially result in the loss of valuable dialect-specific information and hinder the accuracy of the classification process. Finally, the text was tokenized by MARBERT and MARBERTv2 tokenizer utilizing the Transformers library.

3 System Description

This study conducted comprehensive experiments to explore various modifications to our baseline models, MARBERT and MARBERTv2 (Abdul-Mageed et al., 2020a), as detailed in Table 1. We maintained a constant batch size of 64 throughout our experiments and conducted 15 epochs, saving the epoch with the best F1 score by using early stopping. The Adam optimizer (Jais et al., 2019) and Cross Entropy Loss (Smith and Johnson, 2022)

Experiment	Description	Test		Dev	
		Accuracy(%)	F1(%)	Accuracy(%)	F1(%)
Exp.1	MARBERT+ adapter (LR=2e-5)	74.78	74.63	76.06	75.76
Exp.2	MARBERTv2+ adapter (LR=2e-5)	73.58	73.35	75.20	75.01
Exp.3	MARBERT(LR=2e-5)	79.86	80.03	81.61	81.86
Exp.4	MARBERTV2(LR=2e-5)	79.06	79.14	81.11	81.18
Exp.5	MARBERT (last 4 Layers Conc.)(LR=2e-5)	78.39	78.36	79.94	79.87
Exp.6	MARBERTv2(last 4 Layers Conc.) (LR=2e-5)	80.28	80.33	82.44	82.56
Exp.7	MARBERT (average layers 4-7 and conc. output with last 4 layers)(LR=2e-5)	79.86	80.03	80.61	80.72
Exp.8	MARBERTv2 ((average layers 4-7 and conc. output with last 4 layers) (LR=2e-5)	80.83	80.94	81.61	81.86
Exp.9	Repeat Exp.7 + utilizing 1 Conv. Filter(kernel size=5) + MP (LR=2e-5)	81.50	80.83	81.83	81.91
Exp.10	Repeat Exp.8+ utilizing 1 Conv. Filter(kernel size=5)+ MP (LR=2e-5)	81.47	81.43	82.56	82.51
Exp.11	Repeat Exp.7 + BILSTM as classifier (LR=2e-5)	77.72	77.84	78.44	78.33
Exp.12	Repeat Exp.7 +BILSTM + 1 Conv. Filter(kernel size=5) + MP (LR=2e-5)	78.36	78.49	79.11	79.30
Exp.13	Repeat Exp.7 +3 Conv. Filters: kernel sizes(5,4,3) + MP (LR=1e-5)	79.86	80.00	81.83	81.91
Exp.14	Repeat Exp.7 +3 Conv. Filters:kernel sizes(10,8,6) + MP (LR=1e-5)	81.56	81.67	83.06	83.14
Exp.15	Repeat Exp.7 +3 Conv. Filters: kernel sizes(7,7,7) + MP (LR=1e-5)	81.64	81.64	82.72	82.84
Exp.16	Repeat Exp.7 +3 Conv. Filters:kernel sizes(12,10,8) + MP (LR=1e-5)	80.72	80.83	81.61	81.86
Exp.17	Voting Ensemble(Exp 3-16)	83.67	83.73	85.20	85.27
Exp.18	Average Ensemble(Exp 3-16)	83.31	83.36	84.11	84.16

Table 1: Experimental Results for Different Models on Test and Dev Datasets(Abbreviations: F1 - F1-score, MP - Maxpooling, Conc. - Concatenate, Conv. - Convolution)

were utilized in all cases. Learning rates (LR) varied by experimental setup between 1e-5 and 2e-5. Let us delve into more details about each of the 18 different experiments (Exp.) and their significance within this study:

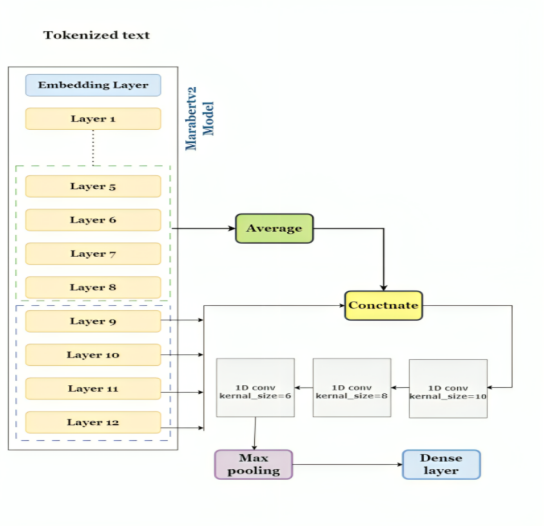

Figure 2: Best Model architecture

3.1 Adapters for MARBERT and MARBERTv2 (Exp.1 and Exp.2)

These experiments explored the impact of adding adapters to the baseline models, MARBERT and MARBERTv2. Adapters are specialized neural network modules added to the models to fine-tune their performance for specific tasks (Pfeiffer et al., 2020) .

3.2 Layer Concatenation (Exp.5 and Exp.6)

In these experiments, the study investigated the concatenation of the last four layers of BERT-Base (Devlin et al., 2018), MARBERT and MARBERT v2 (Abdul-Mageed et al., 2020a). This approach aimed to capture and combine features from different model layers, potentially improving its representation learning capabilities.

3.3 Average Layer 4-7 and Concatenation (Exp.7 and Exp.8)

Experiments 7 and 8 focused on taking an average of layers 4-7 and concatenating it with the last four layers of the models, MARBERT and MARBERTv2, respectively. This approach aimed to leverage layer stacking for enhanced model performance. The results provide insights into the combined impact of these modifications for each model.

3.4 Convolutional Layers with Varying Kernel Sizes (Exp.9 to Exp.16)

These experiments introduced leveraging a series of convolutional layers with varying filter sizes. The ReLU activation function was used within these convolutional layers to introduce non-linearity and enhance the model's capacity to learn complex representations. Following the convolutional layers, max-pooling layers (MaxPool1D) were utilized to reduce the spatial dimensions of the feature maps. The size of the pooling window was determined dynamically based on the length of the convolutional filter. Specifically, the filter size of the last convolutional layer was subtracted from the sequence length, and the result was then added to the stride value. The outputs of these convolutional and max-pooling layers were then flattened. Subsequently, a fully connected dense layer was employed to process the sentence embedding further. (Jacovi et al., 2018).

3.5 Bidirectional LSTM (Exp.11)

Experiment 11 involved adding Bidirectional Long Short-Term Memory (BILSTM) layers as a classifier layer for the MARBERTv2 model. BILSTM layers process input sequences in both forward and backward directions, potentially capturing dependencies in the data more effectively (Graves et al., 2005).

3.6 Ensemble Methods (Exp.17 and Exp.18)

These experiments leveraged ensemble methods to enhance model performance further (Re and Valentini, 2012). The Voting Ensemble (Exp.17) and Average Ensemble (Exp.18) combine the outputs of multiple experiments (Exp.3 to Exp.16) to make predictions. The Voting Ensemble considers the majority or weighted votes, while the Average Ensemble computes the mean of probabilities for predictions.

4 Results and discussion

We present a summary of our experimentation and evaluation of various model enhancements, as

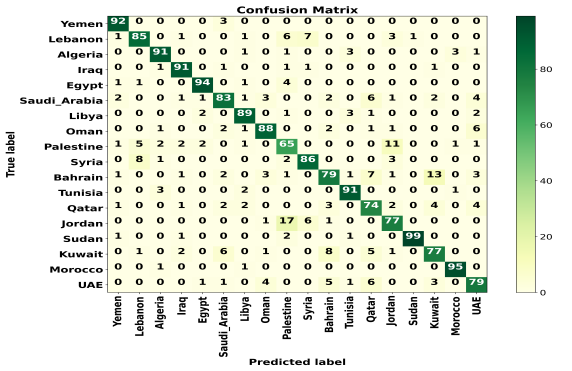

Figure 3: Confusion Matrix for DEV set: voting Ensemble method

reported in Table 1. With a particular emphasis on the F1-score, each experiment assesses the performance achieved by adapting and modifying the baseline models, MARBERT and MARBERTv2.

In our extensive series of experiments, we conduct an exploration of various model enhancements with a primary focus on optimizing F1 scores. Among these experiments, Exp.14 showcases the best results as a standalone model. As shown in Figure 2, this model is built upon the foundation of Exp.7 with the addition of three convolution filters (kernel sizes: 10, 8, 6), followed by max-pooling, and also demonstrates robustness with an impressive F1-score of 81.67% on the test dataset. These results emphasize the significance of spatial feature extraction in text classification tasks. Regarding our methodological approach, Exp.17 represents the most effective method. It serves as our submission and leverages ensemble techniques to combine the predictions of multiple models. This ensemble method significantly outperforms individual models, achieving outstanding F1-scores of 85.27% for the DEV dataset and 83.73% for the test dataset.

Notably, we observe instances of misclassification between the two classes, notably between Jordan and Palestine, as well as between Kuwait and Bahrain, as illustrated in Figure 3. These misclassifications can be attributed to several factors, including historical, cultural, and linguistic nuances that may pose challenges for natural language processing models. The misclassification of content related to Kuwait and Bahrain is a result of shared geographical proximity and cultural ties, leading to overlapping themes and terminology in text data. These overlapping characteristics can cause our models to occasionally struggle in correctly differentiating between the two, resulting in fluctuations in classification performance. These observed misclassifications underscore the need for continued research and model refinement, especially when dealing with regions or topics characterized by subtle distinctions. Addressing such complexities will contribute to improving the accuracy and robustness of models in handling cases with inherent challenges like those presented by Jordan vs. Palestine and Kuwait vs. Bahrain.

With more time available, we will delve into training on larger datasets. Additionally, our study will explore the use of different loss functions for various hyperparameters and incorporate additional ensemble methods such as stacking, bagging, boosting, random forests, AdaBoost, and gradient boosting.

5 Conclusion

Overall, This paper outlines our methods for solving Nuanced Arabic Dialect Identification (NADI) shared task 2023 subtask-1. The extensive experimentation and analysis highlighted the nuanced nature of model enhancements and adaptations. Some modifications, like layer concatenation and the addition of convolution layers, exhibited clear benefits. On the other hand, adapters had more limited impacts. Additionally, ensemble methods emerged as a powerful tool for boosting the score. These findings emphasize the need for a thoughtful and data-driven approach when fine-tuning models for specific tasks and domains in natural language processing. Our system ranks in the 6th best spots of the leaderboards of the first subtask with an F1-score of 83.73%. Future research directions include investigating the impact of larger training datasets on model performance.

6 Limitations

We focused on MARBERT and MARBERTv2 models without comparing them to alternative models. Furthermore, we should have leveraged the advantages of more extensive datasets and various hyperparameters. However, a significant strength of our study lies in exploring the integration of transformers with deep-learning models and adapters.

References

Reem Abdel-Salam. 2022. Dialect & sentiment identification in nuanced arabic tweets using an ensemble of prompt-based, fine-tuned, and multitask bert-based models. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 452–457.

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, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. Nadi 2020: The first nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2010.11334*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

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. *arXiv preprint arXiv:2103.01065*.

Joseph Attieh and Fadi Hassan. 2022. Arabic dialect identification and sentiment classification using transformer-based models. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 485–490.

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

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.

Ron Darvin. 2016. Language and identity in the digital age. *The Routledge handbook of language and identity*, pages 523–540.

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*.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037*.

Imran Khan Mohd Jais, Amelia Ritahani Ismail, and Syed Qamrun Nisa. 2019. Adam optimization algorithm for wide and deep neural network. *Knowledge Engineering and Data Science*, 2(1):41–46.

Abir Messaoudi, Chayma Fourati, Hatem Haddad, and Moez BenHajhmida. 2022. icompass working notes for the nuanced arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 415–419.

Hamada Nayel, Ahmed Hassan, Mahmoud Sobhi, and Ahmed El-Sawy. 2021. Machine learning-based approach for arabic dialect identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 287–290.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Matteo Re and Giorgio Valentini. 2012. Ensemble methods. *Advances in machine learning and data mining for astronomy*, pages 563–593.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic dialects in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 35–40.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Wael Salloum. 2018. *Machine Translation of Arabic Dialects*. Columbia University.

John Smith and Lisa Johnson. 2022. Categorical cross entropy loss for multi-class classification. *Journal of Machine Learning Research*, 30(1):100–120.

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