SANA at NADI 2023 shared task: Ensemble of Layer-Wise BERT-based models for Dialectal Arabic Identification

Nada Almarwani, and Samah Aloufi
Dep. of Computer Science, College of Computer Science and Engineering, Taibah University
nmarwani@taibahu.edu.sa, slhebi@taibahu.edu.sa

Abstract

Our system, submitted to the Nuanced Arabic Dialect Identification (NADI-23), tackles the first sub-task: Closed Country-level dialect identification. In this work, we propose a model that is based on an ensemble of layer-wise fine-tuned BERT-based models. The proposed model ranked fourth out of sixteen submissions, with an F1-macro score of 85.43.

1 Introduction

Arabic is the national language of 25 countries spoken by more than 372 million speakers[1]. While Modern Standard Arabic (MSA) is the formal written language and is used in speech in a formal contexts such as in academia, official communications, and education (Althobaiti, 2020), each country has its own culturally-based dialect that is used in daily communication and informal situations (Elnagar et al., 2021). Nowadays, within the evolution of virtual communication technologies and the intense popularity of social media platforms, dialectal Arabic has replaced MSA as the primary written form of Arabic to generate online informal content. For example, users on social media share news, comment on political and social events, and express opinions concerning various aspects of life using their own dialect. Accordingly, social media is an invaluable resource for harvesting huge amounts of dialectal Arabic data which can be utilized in numerous computational linguistics and Natural Language Processing (NLP) applications. Due to variations between dialects in term of vocabulary usage, meaning, and sense of given words or phrase, automatic identification between unique dialects is a crucial component for improving several downstream applications such as sentiment analysis, speech recognition, and machine translation.

In order to increase the efficiency of Arabic NLP, the Nuanced Arabic Dialect Identification (NADI) shared task series are dedicated to developing solutions for Arabic dialects identification and other related dialectal processes (Abdul-Mageed et al., 2020, 2021b, 2022, 2023). The majority of the works submitted to the NADI-22 employed pre-trained BERT-based models that are specifically trained on Arabic corpus, such as MARBERT (Abdul-Mageed et al., 2021a), ArabBERT (Antoun et al., 2020), and AraGPT2 (Antoun et al., 2021) using various tuning and data augmentation techniques (Abdel-Salam, 2022; Shammary et al., 2022). Other researchers, such as (AlShenaifi and Azmi, 2022) and (Sobhy et al., 2022), used classical machine learning algorithms with TF-IDF and word embeddings. In this paper, following the first line of work, we present our system submitted to the NADI-2023 shared task (Abdul-Mageed et al., 2023). Specifically, to address the first shared sub-task, our approach is based on an ensemble of layer-wise BERT-based models. Each model is trained independently by accessing hidden states from a designated BERT layer and averaging them to generate the final text embeddings.

This paper is organized as follows: Section 2 presents the dataset utilized in our work, Section 3 introduces the proposed system for Arabic dialect identification, Section 4 provides details experimental results and evaluation, Section 5 discusses the model's results and analyze its errors, and finally, Section 6 summarizes findings and possible future work.

[1] https://lingua.edu/the-most-spoken-languages-in-the-world/

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 625–630

Model	Freeze Embeddings	Fine Tuned layers
layer 1	0.760	0.736
layer 2	0.799	0.778
layer 3	0.807	0.795
layer 4	0.819	0.799
layer 5	0.824	0.799
layer 6	0.824	0.803
layer 7	0.835	0.827
layer 8	0.841	0.826
layer 9	0.844	0.830
layer 10	**0.855**	**0.840**
layer 11	0.844	0.839

Table 1: The F1-score macro metrics that were computed independently for each layer-wise model on the development set.

Ensemble Model	Freeze Embeddings	Fine Tuned layers
layers(1-11)	0.865	0.851
layers(2-11)	0.865	0.853
layers(3-11)	0.867	0.856
layers(4-11)	0.866	0.856
layers(5-11)	0.870	0.854
layers(6-11)	**0.874**	0.857
layers(7-11)	0.870	0.857
layers(8-11)	0.872	0.850
layers(9-11)	0.870	0.850
layers(10-11)	0.865	0.853
layer(11)	0.844	0.839

Table 2: The results of F1-score macro metrics on the development set for our ablation study, which is based on an ensemble of the layer-based models.

2 Dataset

The NADI-2023 Shared Task provided the TWT-23 dataset for the Arabic dialects identification task. The dataset contained a total of 23,400 tweets that included 18 Arabic dialects. The dataset was categorized into 18K tweets for training, 1800 tweets for development, and 3600 samples for testing. The training set contained 1000 samples for each dialect class, and the development set included 100 samples for each target class.

3 System Description

Interpretability of pre-trained language models is an outstanding and active research area in NLP. Various studies have been proposed including studies that investigate and analyze the model's implicit representations across intermediate layers (Kakouros and O'Mahony, 2023; Song et al., 2022). Motivated by this line of work, in this paper, we explore the potential of the MARBERTv2 model (Abdul-Mageed et al., 2021a)[2], on Country-level dialects identification task. It should be noted that we also tested other Arabic pre-trained models, such as AraBERT; however, we achieved the best results using MARBERTv2.

Specifically, during the training phase, we fine-tuned 12 independent models based on MARBERTv2. For each model, we chose a specific layer and averaged its hidden states to generate the text embeddings, which then fed through task-specific linear classifier to make the final prediction. Furthermore, we experimented with the model parameters to identify which one to freeze during the fine-tuning, in which the optimal results were obtained by freezing the embeddings layer. During the validation phase, we used a soft voting ensemble method and an ablation study, which we will detail in Section 4.1, to determine the best model. Hence, our final submission was an ensemble of models from layers 6 to 11.

[2]Arabic-based pre-trained BERT model that is publicly available in the HuggingFace library (Wolf et al., 2020)

Rank	Team	F1-Score	Accuracy
1	NLPeople	87.27	87.22
2	rematchka	86.18	86.17
3	Arabitools	85.86	85.81
4	Our team	85.43	85.39

Table 3: Performance of the submitted systems on the leaderboard of sub-task1

Experimental setup We mainly followed the same experimental setups used in (Abdel-Salam, 2022) to fine-tune the model with the exception of the learning rate, weight decay and sentence length, which was set to 2e-5, 1e-2, and 512, respectively. We trained the model with a batch size of 8, for 10 epochs. After each epoch, the model was evaluated on the development set, and the best performant parameters were saved.

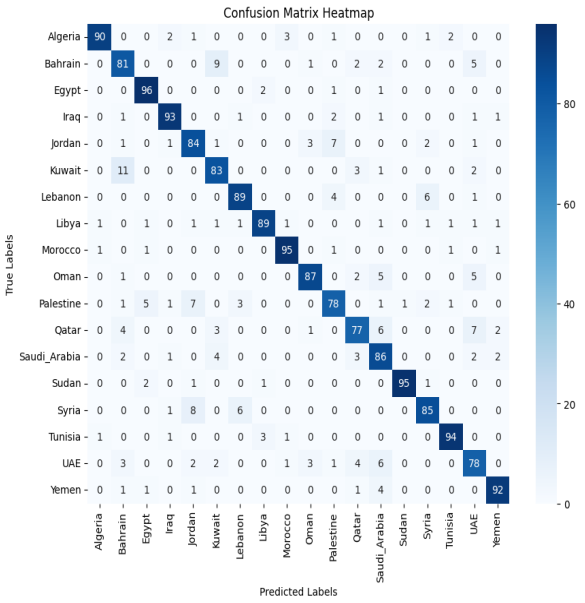

Figure 1: Confusion Matrix Heat-map for development set classification.

Class	Precision	Recall	F1
Algeria	0.97	0.90	0.93
Bahrain	0.76	0.81	0.79
Egypt	0.91	0.96	0.93
Iraq	0.93	0.93	0.93
Jordan	0.80	0.84	0.82
Kuwait	0.81	0.83	0.82
Lebanon	0.89	0.89	0.89
Libya	0.94	0.89	0.91
Morocco	0.94	0.95	0.95
Oman	0.92	0.87	0.89
Palestine	0.82	0.78	0.80
Qatar	0.84	0.77	0.80
KSA	0.75	0.86	0.80
Sudan	0.99	0.95	0.97
Syria	0.87	0.85	0.86
Tunisia	0.95	0.94	0.94
UAE	0.76	0.78	0.77
Yemen	0.93	0.92	0.92

Table 4: F1-score, recall, and precision breakdown of how well the model performs for each individual class.

4 Results

We evaluated the performance of our proposed method on the dialects identification task through a set of experiments to investigate the impact of each layer on classifying the 18 dialects. Table 1 demonstrates the performance of the layer-wise-based models on the development set. The performance improved with the higher layers. Notably, freezing the embeddings during the fine-tuning yielded a better overall performance. Also, averaging the corresponding hidden states of layer 10 while freezing the embeddings achieved the best result with an F1-score of 85.5%. In addition to exploring the layer-wise models' performance independently, we used a soft voting ensemble technique along with an ablation study to select the combination of independent models that yield the best performance on the development set.

4.1 Ablation Study Result

The goal of an ablation study is to examine the impact of removing components of an Artificial intelligence-based system on the system's performance (Zschech, 2022). We examine the impact of different layer-based models on the final model's performance using a soft voting ensemble, as shown in Table 2. Combining models trained on layers 1-11 results in the worst performance; however, removing lower-layer-trained models improved the results. Also, the performance slightly decreased when using models trained only on higher layers (8, 9, 10, and 11). The best results were obtained with an ensemble of models that trained on layers 6 through 11, with an F1-score of 0.874 when embeddings were frozen and 0.857 when embeddings were included in the fine-tuning.

Testing Phase: Table 3 shows the performance of our system submitted to the NADI-2023 shared task: closed country-level dialect identification compared to the top 3 systems.

5 Error Analysis and Discussion

Table 4 shows a detailed evaluation of the model's performance across the 18 distinct classes. Precision values are relatively high at 0.80 for most of the classes. This indicates a strong overall performance, except for the KSA and UAE dialects, where the precision falls under 0.80. Conversely, recall values have less variation. The Algeria, Egypt, Iraq, Morocco, Sudan, and Tunisia classes have high recall rate, which reflects the models' abilities to capture instances from these classes. The F1-score results show the model's strong per-

True Label	Predicted Label	Text	English
Egypt	KSA	لا ياصاحبي مع نفسك هه	No, my friend, with yourself
Egypt	Libya	فيه راجل ينور وشك وفيه راجل يطفيه	There is a man who lights up your face, and there is a man who extinguishes it
Oman	KSA	عطوها قهوه مره	Give her bitter coffee
Yemen	KSA	يمكن النت عندك بطي والا قد نشرتها	Maybe your internet is slow or else you have already posted it
Palestine	KSA	حطي بودره ومي وكم نقطه من ماء الزهر وحطيه ع وجهك لينشف وبعد مابنشف كتيه وغسليه بمي فاتره وشكرا باي عفوا	Apply some powder and water, and a few drops of blossom water on your face. Let it dry, and once it dries, peel it off and wash with water. Thank you, bye, you're welcome
Qatar	Kuwait	يارب بكرا انصدم من سهوله الامتحان واطلع مستانسه وحاله عدل	Dear God, I hope that tomorrow I'll be surprised by how easy the exam is, and I'll come out happy and in a good mood.
Oman	UAE	شو دخل هذا في هذاا	What does this have to do with that
Libya	Morocco	افهمها وهي طايره تويتر راه مش فيس باش نكتب ع راحتي	I understand that this is Twitter not Facebook to write my mind.
Syria	Lebanon	الصغير بدو و الكبير بدو وما حدا عاجبو حالو	The young one wants, and the old one wants, and no one is pleased.
Jordan	Palestine	طيب دعمل حالي مكيفه ع الدوام	Alright, I'll pretend to be cool at work.

Table 5: Examples of Incorrect Predictions from the Development Set.

formance, with most of the classes achieving score of 0.80 or higher.

Figure 1 shows a heat-map of confusion matrix for the development set to further analyze the margin of error in the model's predictions. In general, with minor exceptions, the model seems to perform well for most of the classes. For example, the model preforms well at predicting instances for Egypt, Morocco, and Sudan classes, with true positive exceeding 95 instances. Conversely, the number of true positives are as low as 79 instances or less when predicting instances for the Palestine, UAE, and Qatar classes.

To further analyse, Table 5 shows examples from the development set that our model failed to predict correctly. We observed that the errors of the models of False Positive (FP) and False Negative (FN) fall in one of the following categories:

Missing of diacritics: In Arabic, while different Arabic dialects share common linguistic features, differences remain in the usage of the vocabulary and its meaning. Diacritics plays a crucial role in disambiguate the senses, meanings, and semantics of Arabic language (Matrane et al., 2023; Almuqren and Cristea, 2016; Azmi and Almajed, 2015). We hypothesize that adding diacritics may improve the model's performance in predicting the dialect of a given text. To illustrate more, the first two examples in Table 5 presents this case of ambiguity which might be resolved by diacritics. As can be seen from the confusion matrix, the Egypt class has the least number of FN. We noted that correctly classifying these examples is challenging, even for humans, using the written text only without any context. However, for example, adding diacritics to the word "صاحبي SAHby", which translate in English to "My friend", might help the model to identify the correct class. In particular, in the Egyptian dialect this word would be pronounced with the following diacritics "صاحَبَي SAHabayi", where in the KSA dialect it would be pronounced with the following diacritics "صاحِبَي SAHibayi". Including diacritics may also resolve the ambiguity in the second example, where the words "راجل rAjl and ينور ynwr" in the example, which translate respectively to "man" and "lights up", pronounced differently in both Egyptian and Libyan dialects.

Regional Varieties: Among the 18 dialects classes, the KSA class has the largest variety of dialects due to the geographical diversity and historical migration of people from different linguistic backgrounds. Thus, the East region of KSA tends to share a lot of linguistic similarities with Egypt, while the Southern region share similarities with Yemen, the Northern region is similar to the Levantine dialect (this includes: Syria, Jordan, Palestine, and Lebanon), and the Middle and Western regions congruent with rest of Gulf countries (Bayazed et al., 2020). Also, according to (Alruily, 2020), the majority of most active twitter users are from KSA. Hence, we believe that these factors affected the performance of our model, as the majority of the FP predictions were a result of flawed prediction where other classes were categorized as KSA, examples $3-5$ in Table 5.

Dialects Family: We noted that most of the FP and FN between classes occur among dialects that belong to the same family, or regional varieties of a given dialect. For example, many of the FP and FN occurred in the Gulf dialects family, which includes UAE, Qatar, Bahrain, Kuwait, Oman, Iraq, and certain parts of KSA. This also evident in examples from the Levantine and North African dialects family, example $6-10$ in Table 5.

6 Conclusion

This work describes our proposed system to automatically identifying dialectal Arabic, which has been submitted to the NADI-2023 shared task. The proposed system leveraged the intermediate layers of the pre-trained MARBERTv2 in identifying the Arabic dialects instead of relying on the final layer for text representation. The proposed layer-wise BERT-based models demonstrate a strong overall performance in distinguishing 18 Arabic dialects, achieving an F1 score of 87% on the development set and 85% on the test set. Furthermore, we analyzed the performance of our model and discuss the factors that caused FP and FN predictions. Hence, further elaboration could be followed to study the impact of using diacritics on model performance.

References

Reem Abdel-Salam. 2022. Dialect & sentiment identification in nuanced Arabic tweets using an ensemble of prompt-based, fine-tuned, and multi-task BERT-based models. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 452–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. 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, Online. Association for Computational Linguistics.

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. 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. 2021b. 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.

Latifah Almuqren and Alexandra I Cristea. 2016. Framework for sentiment analysis of arabic text. In *Proceedings of the 27th ACM conference on hypertext and social media*, pages 315–317.

Meshrif Alruily. 2020. Issues of dialectal saudi twitter corpus. *Int. Arab J. Inf. Technol.*, 17(3):367–374.

Nouf AlShenaifi and Aqil Azmi. 2022. Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 464–467, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Aqil M Azmi and Reham S Almajed. 2015. A survey of automatic arabic diacritization techniques. *Natural Language Engineering*, 21(3):477–495.

Afnan Bayazed, Ola Torabah, Redha AlSulami, Dimah Alahmadi, Amal Babour, and Kawther Saeedi. 2020. Sdct: Multi-dialects corpus classification for saudi tweets. *International Journal of Advanced Computer Science and Applications*, 11(11).

Ashraf Elnagar, Sane M. Yagi, Ali Bou Nassif, Ismail Shahin, and Said A. Salloum. 2021. Systematic literature review of dialectal arabic: Identification and detection. *IEEE Access*, 9:31010–31042.

Sofoklis Kakouros and Johannah O'Mahony. 2023. What does bert learn about prosody? In *20th International Congress of Phonetic Sciences ICPhS*. International Phonetics Association.

Yassir Matrane, Faouzia Benabbou, and Nawal Sael. 2023. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, page 101570.

Fouad Shammary, Yiyi Chen, Zsolt T Kardkovacs, Mehwish Alam, and Haithem Afli. 2022. TF-IDF or transformers for Arabic dialect identification? ITFLOWS participation in the NADI 2022 shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 420–424, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mahmoud Sobhy, Ahmed H. Abu El-Atta, Ahmed A. El-Sawy, and Hamada Nayel. 2022. Word representation models for Arabic dialect identification. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 474–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mingyang Song, Yi Feng, and Liping Jing. 2022. Utilizing bert intermediate layers for unsupervised keyphrase extraction. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 277–281.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Patrick Zschech. 2022. Beyond descriptive taxonomies in data analytics: A systematic evaluation approach for data-driven method pipelines. *Inf. Syst. E-Bus. Manag.*, 21(1):193–227.