Building an Ensemble of Transformer Models for Arabic Dialect Classification and Sentiment Analysis

Abdullah Khered[1,2] , **Ingy Abdelhalim**[1] and **Riza Batista-Navarro**[1]
[1]The University of Manchester, UK
[2]King Abdulaziz University, Saudi Arabia
abdullah.khered@postgrad.manchester.ac.uk
ingy.abdelhalim@student.manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

In this paper, we describe the approaches we developed for the Nuanced Arabic Dialect Identification (NADI) 2022 shared task, which consists of two subtasks: the identification of country-level Arabic dialects and sentiment analysis. Our team, UniManc, developed approaches to the two subtasks which are underpinned by the same model: a pre-trained MARBERT language model. For Subtask 1, we applied undersampling to create versions of the training data with a balanced distribution across classes. For Subtask 2, we further trained the original MARBERT model for the masked language modelling objective using a NADI-provided dataset of unlabelled Arabic tweets. For each of the subtasks, a MARBERT model was fine-tuned for sequence classification, using different values for hyperparameters such as seed and learning rate. This resulted in multiple model variants, which formed the basis of an ensemble model for each subtask. Based on the official NADI evaluation, our ensemble model obtained a macro-F1-score of 26.863, ranking second overall in the first subtask. In the second subtask, our ensemble model also ranked second, obtaining a macro-F1-PN score (macro-averaged F1-score over the `Positive` and `Negative` classes) of 73.544.

1 Introduction

There are approximately 400 million Arabic speakers worldwide, spread geographically in 22 countries around the world (Boudjellal et al., 2021). With early manifestations of Arabic dating back to the 8th century BCE, the Arabic language has been redefined and refined over many decades across different continents. Many scholars struggled to define Arabic as a single language, with many considering Classical Arabic (CA)—the language of the Quran—as the ideal archetype. In modern times, Modern Standard Arabic (MSA) has been used in most official publications, broadcasts, political speeches, and written texts. However, most people use spoken varieties of Arabic in their daily lives. Some of these spoken varieties differ from each other significantly and are almost mutually unintelligible, whilst others bear strong similarities. These spoken variations of Arabic are commonly referred to as Dialectical Arabic (DA).

Thus far, the majority of the research in Arabic Natural Language Processing (NLP) has overlooked the variations across the different Arabic dialects (Oueslati et al., 2020), largely due to the lack of datasets that take the different DA types into consideration. The goal of the Nuanced Arabic Dialect Identification (NADI) shared task series is to diminish this research gap, by providing datasets where examples are organised according to dialects (Abdul-Mageed et al., 2020, 2021b, 2022). As part of the NADI 2022 shared task, organisers made available datasets that support two sub-tasks, namely, dialect identification (Subtask 1) and sentiment analysis of country-level dialectical Arabic (Subtask 2).

Recent advancements in NLP research have led to the development of transformer-based language models which learn contextual embedding representations of sequences, and which have been shown to obtain state-of-the-art performance on many NLP tasks (Vaswani et al., 2017; Liu et al., 2020; Nagoudi et al., 2022). MARBERT (Abdul-Mageed et al., 2021a) is a language model that was pre-trained specifically on DA, and formed the basis of our approach to the NADI 2022 shared tasks.

2 Datasets

NADI 2022 is the third in the NADI shared tasks series and consists of two subtasks. Similar to past editions of the shared task, the first subtask is a multi-class classification problem aimed at recognising the Arabic dialects used in tweets. Unlike in previous years, however, a new task focussing on sentiment analysis of dialectical Arabic tweets was

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 479 - 484
December 8, 2022 ©2022 Association for Computational Linguistics

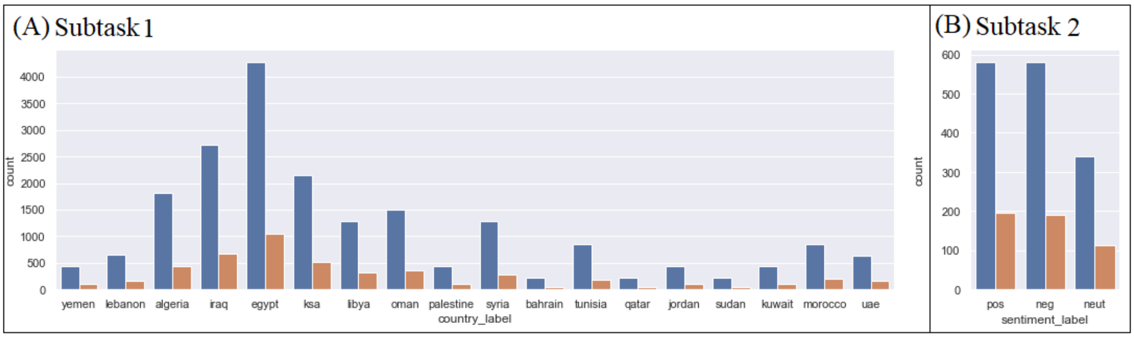


Figure 1: Label distribution of the training and the development sets for Dialect Arabic Identification (Subtask 1) and Sentiment Analysis (Subtask 2).

organised and put forward as the second subtask.

The organisers prepared a dataset of labelled tweets covering 18 Arab countries for the dialect identification subtask. It was split into training, development, and two test sets. Whilst the first test set (Test-A) covers 18 country-level dialects (as the training and development sets do), the second one (Test-B) includes an unknown number of dialects.

The distribution of examples across the different classes of interest for each of the subtasks is shown in Figure 1. As one can observe in Figure 1-A, the distribution across the 18 dialects is unbalanced, with Eqypt being the most frequently occurring label in the dataset for Subtask 1.

For the sentiment analysis subtask, the organisers provided a dataset of tweets labelled as any one of three classes: Positive, Negative and Neutral. It was divided into training, development and test sets. As shown in Figure 1-B, the Positive and Negative classes have an almost equal distribution between them, but the Neutral class has a slightly lower number of training samples.

The datasets for both subtasks were preprocessed whereby URLs were replaced with the token 'URL', and Twitter usernames were replaced with the token 'USER', in order to normalise them.

3 Methodology

Our approaches to the two subtasks are both underpinned by the first version of MARBERT, a language model that had been trained on a 128GB dataset containing both MSA and DA tweets (Abdul-Mageed et al., 2021a).

It is worth noting that we built our own version of the MARBERT model by continuing to train it for the masked language modelling (MLM) objective (Devlin et al., 2019); we describe this model in detail in Section 3.2 below. However, our experiments showed that using our own MARBERT model led to performance improvement only for sentiment analysis and not for dialect identification. Therefore this model formed the basis of our solution for Subtask 2 but not for Subtask 1.

3.1 Subtask 1: Dialect Identification

The original pre-trained MARBERT model was fine-tuned for dialect identification using the full training set for Subtask 1 that was provided by the NADI organisers. Considering the imbalance in the distribution of training samples across the different classes (as shown in Figure 1), it was unsurprising that when evaluated on the development set, the resulting sequence classification model is unable to predict the least represented classes (e.g., Bahrain and Qatar), but obtains satisfactory performance for the classes with sufficient examples.

Therefore, we investigated the use of undersampling, whereby the training samples belonging to the over-represented classes such as Egypt and KSA (Kingdom of Saudi Arabia), were reduced. Our undersampling technique is based on the removal of randomly selected samples (Chawla, 2010) from the over-represented classes; this led to the creation of a version of the dataset where the number of samples for each class was capped at 215 (i.e., the number of samples in the least represented dialects, namely, Bahrain, Qatar and Sudan). However, we also created other dataset versions where the number of samples per class was capped at 250 and 300. In this case, it was necessary to apply oversampling on the least represented classes (Chawla, 2010); to this end, randomly selected samples in those classes were duplicated. Our initial experiments showed that fine-tuning the original MARBERT model on these balanced versions of the dataset led to classification models that are able to predict the least represented dialects, although their performance on the sufficiently represented dialects was degraded compared with a model fine-tuned on the full training set.

Considering that fine-tuning on the full training set and fine-tuning on the balanced data, each has its own advantages, our solution for this subtask was based on combinations of models resulting from both.

3.2 Subtask 2: Sentiment Analysis

Taking the checkpoint for the original pre-trained MARBERT model[1], we continued to train it for masked language modelling using the dataset of 10 million unlabelled Arabic tweets, that was provided by the NADI organisers as part of the shared task. Out of these tweets, 90% were used for training, whilst the remaining 10% were used for validation. Both the number of epochs and batch size were arbitrarily set to 8 and the maximum sequence length was fixed at 512. The resulting model was then fine-tuned for sentiment analysis using the labelled tweets in the training set for Subtask 2. We also considered creating a version of the dataset where the dominant classes, i.e., `Positive` and `Negative`, are undesampled. However, models fine-tuned on this version obtained inferior classification performance. Thus, only models fine-tuned on the full training set comprise our solution for this subtask.

3.3 Hyperparameter Optimisation

For each of the subtasks, we trained a number of model variants using the full training sets for both Subtasks 1 and 2, and additionally, on the balanced versions of the training set for Subtask 1. These model variants are based on the exploration of a range of values for seed and learning rate. Specifically, seed values ranging between 20 and 300 (inclusive) were investigated; we found that setting the seed to 200 led to optimal performance in both subtasks, based on results on the respective development sets. Meanwhile, optimal performance was obtained by setting the learning rate to values ranging between 1.5e^{-5} and 2.5e^{-5} (inclusive).

The batch size was fixed at 32, while the number of epochs was arbitrarily set to 8. For every training run (on Nvidia A100 GPUs), the model trained in the epoch where the best macro-averaged F1-score was obtained, was considered as the best-performing model for that run.

3.4 Ensemble Models

After hyperparameter optimisation, the eight best-performing Subtask 1 models (according to F1-score), were selected: four based on training on the full training set, and the other four based on training on the balanced data. Meanwhile, for Subtask 2, we selected the five best-performing models (based on F1-score) trained on the full training set.

For each subtask, we aimed to identify an ensemble model (Rokach, 2019) that is based on the combination of the predictions of these best-performing models. In Subtask 1, for example, there are 255 possible combinations of the eight models (i.e., $2^8 - 1$ combinations). For each combination (ensemble), the average of the prediction probabilities output by the models for each class was taken as the basis for the overall prediction of the ensemble. A similar process was applied to the 31 possible combinations of the five models for Subtask 2 (i.e., $2^5 - 1$ combinations).

For each of the two subtasks, the three best-performing ensemble models were identified based on experiments on the corresponding development set and formed the basis of our official submission to NADI 2022.

4 Evaluation and Results

The performance of our ensemble models for the dialect identification subtask is summarised in Table 1. Our best-performing model (Ens 1.1) obtained a macro-averaged F1-score of 35.625 on the development set. Meanwhile, the macro-averaged F1-scores on the two test sets are: 34.778 on Test-A (the test set that covers 18 dialects) and 18.948 on Test-B (the test set with an unknown number of dialects). Nevertheless, it is worth noting that our best ensemble model ranks third when evaluated using Test-A, and ranks first when evaluated using Test-B, amongst the submissions from the 19 teams who participated in Subtask 1. If one takes the mean of the macro-averaged F1-scores on Test-A and Test-B as the overall performance for Subtask 1, our best ensemble model ranks second, with a mean score of 26.863.

With regard to the second subtask, we present the performance of our ensemble models for sentiment analysis in Table 3. Instead of the macro-

[1] https://github.com/UBC-NLP/marbert#6-download-arbert-and-marbert-checkpoints

Model	Eval. data	Macro-F1	Acc.
Ens 1.1	Dev	**35.625**	**53.890**
	Test-A	**34.778**	**52.333**
	Test-B	**18.948**	36.839
Ens 1.2	Dev	35.031	53.069
	Test-A	34.152	51.303
	Test-B	17.984	36364
Ens 1.3	Dev	34.937	52.782
	Test-A	34.248	51.366
	Test-B	18.435	**36.974**

Table 1: Results for Subtask 1 based on three different ensemble (Ens) models.

Model	Eval. data	Macro-F1-PN	Acc.
Ens 2.1	Dev	**77.262**	**72.400**
	Test	**73.544**	**67.700**
Ens 2.2	Dev	76.904	**72.400**
	Test	73.200	67.333
Ens 2.3	Dev	76.709	**72.400**
	Test	73.432	67.667

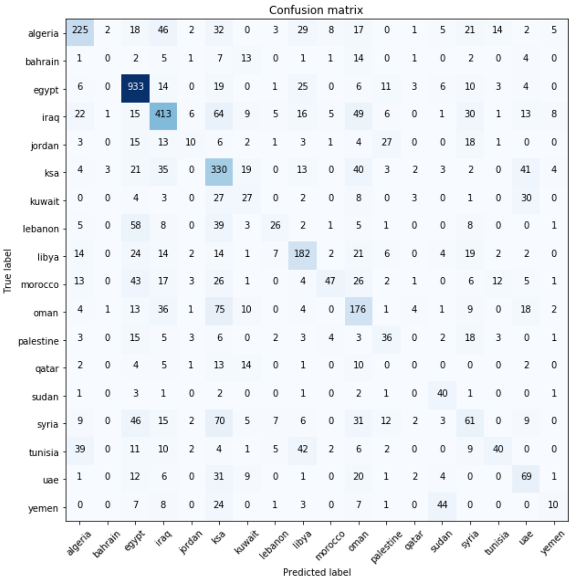
Table 2: Results for Subtask 2 based on three different ensemble (Ens) models.

averaged F1-score over all classes, a different metric (macro-F1-PN) based on the macro-averaged F1-score over the Positive and Negative classes only, was used in the evaluation of this subtask. Our best-performing ensemble model (Ens 2.1) obtained a macro-F1-PN score of 77.262 on the development set and 73.544 on the test set. This model ranks second amongst the submissions from 10 teams who participated in Subtask 2.

5 Discussion

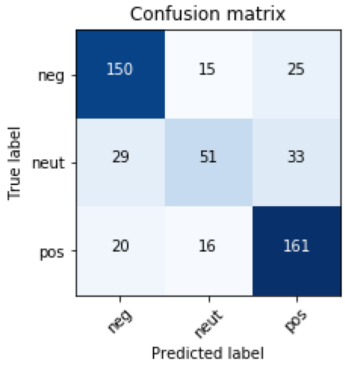
To allow us to draw some insights on the class-level performance of our best-performing dialect identification model, we provide the confusion matrix based on the development set, in Figure 2.

As one can observe in the confusion matrix, the majority of the true samples from dialects such as Egypt, KSA, and Iraq, have been correctly predicted by our model. This can be explained by the fact that such classes are over-represented in the training data. However, the over-representation of such classes is likely to have also led to a detrimental effect, i.e., the model being biased towards such dominant dialects, as can be observed in the columns of the confusion matrix, where many samples tend to be wrongly predicted as Egypt or KSA, for instance. Meanwhile, as expected, the model obtained poor performance with respect to the least represented dialects such as Bahrain and Qatar. Also, our model tends to be confused by dialects which correspond to regions which are geographically close to each other and hence share certain dialects, e.g., Oman vs KSA, Lebanon vs Egypt.



	algeria	bahrain	egypt	iraq	jordan	ksa	kuwait	lebanon	libya	morocco	oman	palestine	qatar	sudan	syria	tunisia	uae	yemen
algeria	225	2	18	46	2	32	0	3	29	8	17	0	1	5	21	14	2	5
bahrain	1	0	2	5	1	7	13	0	1	1	14	0	1	0	2	0	4	0
egypt	6	0	933	14	0	19	0	1	25	0	6	11	3	6	10	3	4	0
iraq	22	1	15	413	6	64	9	5	16	5	49	6	0	1	30	1	13	8
jordan	3	0	15	13	10	6	2	1	3	1	4	27	0	0	18	1	0	0
ksa	4	3	21	35	0	330	19	0	13	0	40	3	2	3	2	0	41	4
kuwait	0	0	4	3	0	27	27	0	2	0	8	0	3	0	1	0	30	0
lebanon	5	0	58	8	0	39	3	26	2	1	5	1	0	0	8	0	0	1
libya	14	0	24	14	2	14	1	7	182	2	21	6	0	4	19	2	2	0
morocco	13	0	43	17	3	26	1	0	4	47	26	2	1	0	6	12	5	1
oman	4	1	13	36	1	75	10	0	4	0	176	1	4	1	9	0	18	2
palestine	3	0	15	5	3	6	0	2	3	4	3	36	0	2	18	3	0	1
qatar	2	0	4	5	1	13	14	0	1	0	10	0	0	0	0	0	2	0
sudan	1	0	3	1	0	2	0	0	1	0	2	1	0	40	1	0	0	1
syria	9	0	46	15	2	70	5	7	6	0	31	12	2	3	61	0	9	0
tunisia	39	0	11	10	2	4	1	5	42	2	6	2	0	0	9	40	0	0
uae	1	0	12	6	0	31	9	0	1	0	20	1	2	4	0	0	69	1
yemen	0	0	7	8	0	24	0	1	3	0	7	1	0	44	0	0	0	10

Figure 2: Confusion matrix for our best-performing dialect identification ensemble model, based on the development set.



	neg	neut	pos
neg	150	15	25
neut	29	51	33
pos	20	16	161

Figure 3: Confusion matrix for our best-performing sentiment analysis ensemble model, based on the development set.

As for our best-performing sentiment analysis model, the confusion matrix in Figure 3 shows that the model performs almost equally well on the Positive and Negative classes. Unsurprisingly, it does not perform as well for the Neutral class, which has a slightly lower number of training samples.

Hypothesising that limited context in any given tweet leads to wrong predictions, we investigated

	Tweet Text	English Translation	Gold	Pred.
1	الله يحفظوا و يحفظنا	May God protect him and protect us	Iraq	Oman
2	ربي يخليهم لك	May god keep them for you	Libya	Oman
3	لا دي مصرية	No, she is Egyptian	KSA	Egypt
4	بس حلو المسلسل	But the series is nice	Jordan	Iraq
5	اي لوف يو	I love you	KSA	Iraq
6	باك من قطر :(	Back from Qatar :(	KSA	UAE
7	الله يحفظه ويطول بعمره	May Allah protect him and prolong his age	KSA	Oman
8	الله يسلمك ، امين	God bless you, amen	KSA	Oman

Table 3: Some of the incorrectly predicted samples, their English translation, their labels in the development set (Gold) and our model's predicted label (Pred).

whether the length of a tweet in terms of number of tokens, has a detrimental impact on model performance. There are 864 samples in the development set with at most four tokens; the macro-averaged F1-score obtained by our model on these samples is 25.180. In contrast, the same model obtained a substantially higher macro-averaged F1-score of 37.385 on the remaining 4007 samples which have four or more tokens. Moreover, as we increased the number of tokens being considered, the model's performance also improved: the macro-averaged F1-score on samples with no more than five tokens (1336 samples) and six tokens (1823 samples) is 26.126 and 28.323, respectively.

Based on the above observations and some samples (that we manually analysed), we argue that defining Arabic dialect identification task as a classification task with a large number of classes (e.g., 18), inevitably leads to overlap. In this scenario, a given tweet could easily qualify as belonging to more than one dialect, where even humans would disagree on the dialect used. This is because many countries may use the same phrase or wording; especially in cases where a tweet contains only a few tokens, it can be extremely hard to pinpoint its country or region of origin.

Table 3 shows some samples from the development set that were wrongly predicted by our model. These samples contain only a few tokens thus making it very challenging to identify their dialect. In fact, some of these samples cannot be identified as one dialect since they can be used in multiple countries. For example, the first four tweets (Samples 1, 2, 3 and 4) in Table 3 were labelled as being from a different dialect to what our model predicted them as; however, they can also be considered as the Egypt or KSA dialects since these phrases are commonly used in Egypt and Saudi Arabia. Moreover, we found samples that include English words, such as Sample 5 which was given KSA and Iraq as its label in the development set and by our model, respectively, when in reality it was not even written in Arabic. It is instead a transliteration of the English phrase *"I love you"*. Similarly, Sample 6 contains the word *"back"* transliterated into Arabic leaving only two Arabic words which translate to *"from Qatar"* from which it is impossible to detect a dialect even by a native Arabic speaker.

We also investigated some samples from neighbouring countries such as KSA, Oman and UAE (United Arab Emirates), which are all Gulf countries. As shown in Table 3, some samples (such as Samples 6, 7 and 8) are not easy to identify since there are some similarities between neighbouring countries' dialects. We thus believe that the task of identifying Arabic dialects could be more suitable as a multi-label classification task whereby each sample can be assigned more than one dialect.

6 Conclusion and Future Work

In this paper, we presented our ensemble-based approaches to the NADI 2022 subtasks: dialect identification and sentiment analysis. Our results demonstrate that an ensemble model consisting of a combination of MARBERT models fine-tuned in different ways, for each of the subtasks, obtains top-ranking performance. A potential future direction is the exploration of multi-task learning for jointly training a model on the two subtasks.

References

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, 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 Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Nada Boudjellal, Huaping Zhang, Asif Khan, Arshad Ahmad, Rashid Naseem, Jianyun Shang, and Lin Dai. 2021. ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. *Complexity*, 2021:1–6.

Nitesh V. Chawla. 2010. *Data Mining for Imbalanced Datasets: An Overview*, pages 875–886. Springer US, Boston, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in arabic language. *Future Generation Computer Systems*, 112:408–430.

L. Rokach. 2019. *Ensemble Learning: Pattern Classification Using Ensemble Methods*. Series in machine perception and artificial intelligence. World Scientific Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.