Ahmed and Khalil at NADI 2022: Transfer Learning and Addressing Class Imbalance for Arabic Dialect Identification and Sentiment Analysis

Ahmed Oumar El-Shangiti
Independent Researcher
Marrakesh, Morocco
ahmedmohamedlemin@gmail.com

Khalil Mrini
Meta AI
Seattle, United States
khalil@meta.com

Abstract

In this paper, we present our findings in the two subtasks of the 2022 NADI shared task. First, in the Arabic dialect identification subtask, we find that there is heavy class imbalance, and propose to address this issue using focal loss. Our experiments with the focusing hyperparameter confirm that focal loss improves performance. Second, in the Arabic tweet sentiment analysis subtask, we deal with a smaller dataset, where text includes both Arabic dialects and Modern Standard Arabic. We propose to use transfer learning from both pre-trained MSA language models and our own model from the first subtask. Our system ranks in the 5th and 7th best spots of the leaderboards of first and second subtasks respectively.

1 Introduction

The 2022 **N**uanced **A**rabic **D**ialect **I**dentification (NADI) shared task (Abdul-Mageed et al., 2022) is comprised of two subtasks: Arabic dialect identification, and sentiment analysis for Arabic dialects. The aim of the shared task is to alleviate the lack of resources in NLP for Arabic dialects, amid growing interest in Arabic dialect language models (Elgezouli et al., 2020; Abdaoui et al., 2021; Issam and Mrini, 2022). The 2022 edition is the third NADI shared task. The 2021 (Abdul-Mageed et al., 2021b) and 2020 NADI shared tasks (Abdul-Mageed et al., 2020) focus on country- and province-level Arabic (sub-)dialect identification. These two editions also tackled tweets in Arabic dialects, gathering dialects from 100 provinces in 21 Arab countries.

In this paper, we tackle both subtasks, using both transfer learning from pre-trained language models, and transfer learning from one subtask to the other, as well as loss functions adapted to the class imbalance in the dataset.

The first subtask tackles country-level Arabic dialect identification in tweets. We first analyse the data, and find that there is a high class imbalance between the 18 countries represented in the tweets. We find that the largest class has nearly 20 times as many samples as the smallest one. We try multiple pre-trained Arabic language models, and find that the highest-performing model is MarBERT (Abdul-Mageed et al., 2021a). We try different loss functions, and find that focal loss (Lin et al., 2017) performs the best, as it applies a modulating term to the cross-entropy loss, enabling the training process to focus on wrongly classified samples. We fine-tune the *focusing* hyperparameter γ, and observe how performance fluctuates accordingly.

The second subtask deals with sentiment analysis for tweets in various Arabic dialects, as well as in Modern Standard Arabic. There are three classes: positive, negative, and neutral sentiment. In our data analysis, we find that there is less class imbalance in the second subtask, especially between the positive and negative classes. However, this second subtask has a much smaller training set, and therefore needs a supplement of knowledge from other sources. Given that external labeled data is not allowed, we decide to employ transfer learning, by fine-tuning the best model from the first subtask on this second one. As the dataset of the second subtask contains both Arabic dialects and Modern Standard Arabic, we hypothesize that performance will benefit from language models trained on Modern Standard Arabic, as well as from data in Arabic dialects. Finally, we show that our system ranks in the 5th and 7th best spots of the leaderboards in the first and second subtasks respectively, and propose suggestions for improving performance.

2 Data

In this section, we describe the data used for training our system in both subtasks.

The first subtask deals with Arabic Dialect Identification. The training data contains 18 classes. Each class corresponds to the national vernacular

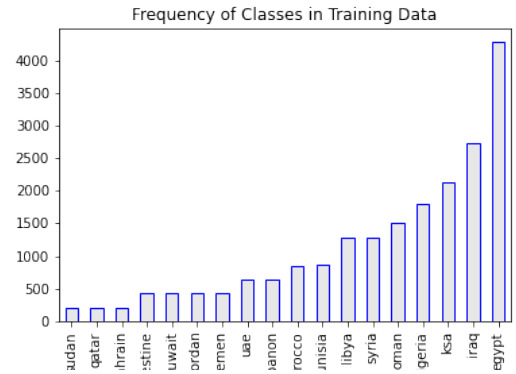

Figure 1: Distribution of country labels for training samples for the first subtask.

Subtask	F1	Acc.	Prec.	Recall
1	0.3305	0.5231	0.3629	0.3411
Subtask	F1	Acc.	Prec.	Recall
2	0.7334	0.6860	0.6658	0.6483

Table 1: Validation set results for our team in both subtasks. Results are computed by the online platform.

of a distinct Arab country. There are 20,398 training samples – all are tweets. We plot the distribution of country labels for training samples in Figure 1. The dataset is unbalanced, as we notice Egypt has 4,283 samples, whereas the smallest classes (Bahrain, Sudan, Qatar) have only 215 samples each.

We perform a similar analysis for validation data, and find that the distribution is similar, as shown in Figure 2. The validation dataset has 4,871 samples. The class with the most samples is again Egypt with 1,041 datapoints, whereas the smallest ones are Qatar and Bahrain with 52 samples each.

The second subtask is Sentiment Analysis over tweets in various Arabic dialects. This is a three-way classification problem, where the goal is to predict whether a tweet – regardless of the arabic dialect – has positive, neutral or negative sentiment. This subtask has fewer datapoints than the first one. The training set contains 1,500 samples, whereas the validation set contains 500 samples. There is roughly the same distribution over the sentiment classes between the two sets, as shown in Figures 3 and 4.

3 System Description

For both subtasks, we investigate the potentials of transfer learning for different Arabic BERT-based models. Specifically, we compared the follow-

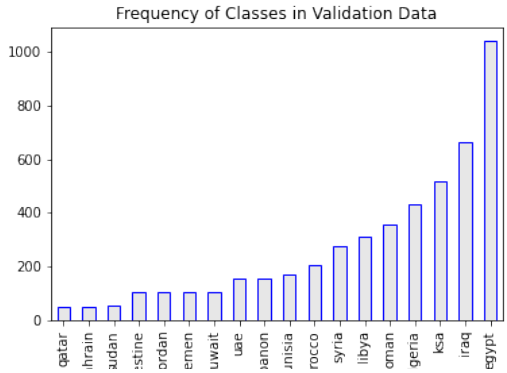

Figure 2: Distribution of country labels for validation samples for the first subtask.

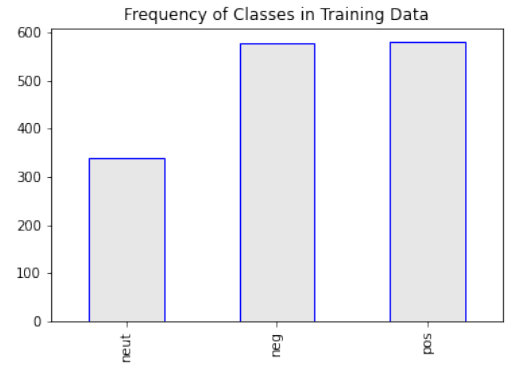

Figure 3: Distribution of sentiment labels for training samples for the second subtask.

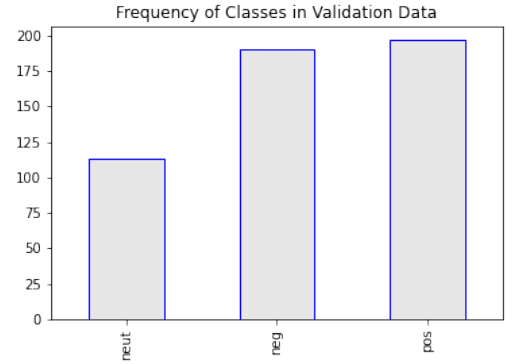

Figure 4: Distribution of sentiment labels for validation samples for the second subtask.

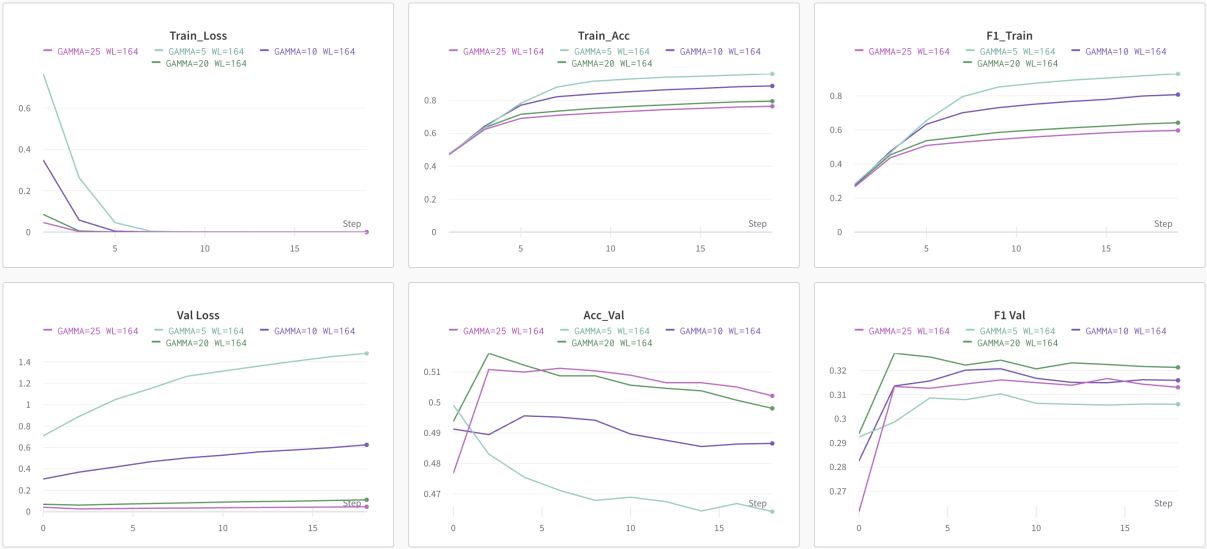

Figure 5: Graphs showing the progression of the loss, accuracy, and F1 scores for the training and validation sets of the first subtask on Arabic Dialect Identification. We change the values of the γ of the Focal Loss, varying them from 5 to 25.

ing pre-trained BERT-based models: MarBERT (Abdul-Mageed et al., 2021a), CamelBERT (Inoue et al., 2021) and AraBERT (Antoun et al., 2020).

Our experiments consist of fine-tuning a pre-trained BERT model, plus one or more fully connected layers. It turns out that the best performance is achieved using only the pre-trained model plus a classification layer.

For all experiments, we use the following hyperparameters: a learning rate of $4 * 10^{-5}$, 10 training epochs, an Adam optimizer with weight decay regularization. The batch size is set to 32 for the first subtask, and 8 for the second subtask.

We implement our models using Pytorch. For the loss functions, we experiment with self-adjusting Dice Loss (`SelfAdjDiceLoss`) (Li et al., 2020), Negative Log-Likelihood Loss (`NLLLoss`), Cross-Entropy Loss (`CrossEntropyLoss`) with and without weighted classes, and Focal Loss (`FocalLoss`) (Lin et al., 2017). The latter has shown the best performance for both sub tasks. This could be due to the fact that the first subtask's dataset is imbalanced, and Focal Loss is designed to alleviate class imbalance. In order to focus on hard, wrongly classified samples, Focal Loss applies a modulating term to the cross-entropy loss. Given the cross-entropy loss formula:

$$\text{CEL}(p_t) = -log(p_t) \tag{1}$$

the focal loss formula is as follows:

$$\text{FL}(p_t) = (1 - p_t)^{\gamma} * [-log\,(p_t)] \tag{2}$$

where γ is the *focusing* hyperparameter. The higher the hyperparameter, the more the focal loss function will focus on wrongly classified samples.

Among the three pre-trained models considered, we found that MarBERT performs the best, in a fair evaluation with fixed hyperparameters. During our experiments, we found that the best configuration is a pre-trained MarBERT model, with a single classification layer, and a Focal Loss function.

Participants of the shared task were not allowed to use external labeled data for training. However, the second subtask has a substantially smaller training set than the first one. We decide to leverage the knowledge learned by the model during the first subtask, and fine-tune the model on the training set of the second subtask.

4 Results and Discussion

For the first subtask, we experiment with the γ hyperparameter of the Focal Loss. We try the following values: 5, 10, 20 (default value), and 25. We show the results on the validation and training sets in Figure 5. We see that the lowest validation loss is achieved with $\gamma = 25$, but the highest accuracy and F1 scores are achieved with $\gamma = 20$. So we use $\gamma = 20$ for the remainder of the experiments. This confirms that performance is higher when class imbalance is addressed during training. The accuracy and F1 scores seem to peak for the

TABLE 4. Leaderboard of Subtask 2

Rank	Team	Macro-F1-PN	Accuracy	Recall	Precision
1	rematchka	75.1555	69.7000	66.2230	67.5684
2	UniManc	73.5443	67.7000	63.9228	65.2702
3	BhamNLP	73.4566	67.3333	62.8315	65.2415
4	pythoneers	73.3959	68.2333	65.8708	66.0751
5	Ahmed_and_Khalil	71.4569	66.0333	63.7342	63.8411
6	giyaseddin	71.4278	65.8000	62.1962	63.5143
7	ISL_AAST	70.5527	64.9667	61.4095	62.5844
8	ANLP-RG	67.3106	61.9000	59.6697	59.6920
9	RUTeam	61.0675	56.1667	53.5776	53.8966
10	Oscar_Garibo	46.4261	43.0000	41.9179	41.9985

Figure 6: All 10 teams in the leaderboard for the second subtask on Sentiment Analysis for Arabic Dialects.

TABLE 1. Leaderboard of Subtask 1

Rank	Team	Average Macro-F1
1	rematchka	27.06
2	UniManc	26.86
3	GOF	26.44
4	mtu_fiz	25.50
5	iCompass	25.32
6	ISL-AAST	24.59
7	Ahmed_and_Khalil	24.35
8	pythoneers	24.12
9	giyaseddin	22.42
10	SQU	22.42

Figure 7: Top 10 teams in the leaderboard for the first subtask on Arabic Dialect Identification.

validation set at the second epoch, whereas they increase for the training set as the training epochs progress. This indicates that overfitting occurs after the second epoch, in particular for $\gamma = 20$.

We show our dev set results in Table 1. For both subtasks, the results are computed by the Codalab online platform based on the predictions of our system. The metrics are macro-F1, accuracy, precision, and recall. Macro-F1 gives equal weight to each class, which matters for the first subtask where there is heavy class imbalance.

For the test set results, our system scores in the 7th best spot in the first subtask, out of 19 participants, and the 5th best spot out of 10 participants in the second subtask. The leaderboards and test results for the first and second subtasks are shown in Figures 7 and 6 respectively. For the first subtask, there are two test subsets: Test-A is a subset containing all 18 classes, whereas Test-B is a subset containing k classes, where k is unknown. The results shown in Figure 7 are the Average of the Macro-F1 scores between both test subsets. We notice that the results of the second through fifth rows in the leaderboard of the first subtask are close. For the second subtask, the shared task organizers evaluate using "*Macro-F1-PN*", which is a Macro-F1 score computed for the Positive and Negative classes, ignoring the Neutral cases.

If we had more time, we would investigate Domain Adversarial learning and Multi-Task Learning. As transfer learning proved useful in the second subtask, this suggests that a multi-task learning setting could benefit both subtasks. Moreover, in the second subtask, there are tweets from different countries, but it is a feature that does not matter in the sentiment analysis task. The model would benefit from learning to not distinguish between Arabic dialects, as it would learn *dialect-agnostic* sentiment features that enable easy knowledge transfer between tweets in different Arabic dialects.

5 Conclusions

In this paper, we presented our team's approach to the two subtasks of the 2022 NADI shared task. We first analysed the data, and find that there is class imbalance between the 18 classes of the Arabic dialect identification subtask. In the Arabic tweet sentiment analysis subtask, we find that classes are relatively more balanced, but there are fewer

datapoints to train on.

We propose to train on MarBERT, and we find that Focal Loss is the loss function that performs best, as it addresses class imbalance. Our experiments with the γ focusing hyperparameter show that we need a large γ value for high F1 scores, confirming that focal loss alleviates class imbalance.

Finally, our system scores favorably in the leaderboards in both subtasks. We suggest for the second subtask that domain-adversarial training could benefit performance, as it would make the model learn dialect-agnostic features about the sentiment classes. We release our code to encourage further research.

References

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

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 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.

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.

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)*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mukhtar Elgezouli, Khalid N Elmadani, and Muhammed Saeed. 2020. Sudabert: A pre-trained encoder representation for sudanese arabic dialect. In *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–4. IEEE.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Abderrahmane Issam and Khalil Mrini. 2022. Goud. ma: a news article dataset for summarization in moroccan darija. In *3rd Workshop on African Natural Language Processing*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.