Arabic Dialect Identification and Sentiment Classification using Transformer-based Models

Joseph Attieh
Huawei Technologies Oy., Finland
joseph.attieh@huawei.com

Fadi Hassan
Huawei Technologies Oy., Finland
fadi.hassan@huawei.com

Abstract

In this paper, we present two deep learning approaches that are based on AraBERT, submitted to the Nuanced Arabic Dialect Identification (NADI) shared task of the Seventh Workshop for Arabic Natural Language Processing (WANLP 2022). NADI consists of two main sub-tasks, mainly country-level dialect and sentiment identification for dialectical Arabic. We present one system per sub-task. The first system is a multi-task learning model that consists of a shared AraBERT encoder with three task-specific classification layers. This model is trained to jointly learn the country-level dialect of the tweet as well as the region-level and area-level dialects. The second system is a distilled model of an ensemble of models trained using K-fold cross-validation. Each model in the ensemble consists of an AraBERT model and a classifier, fine-tuned on (K-1) folds of the training set. Our team Pythoneers achieved rank 6 on the first test set of the first sub-task, rank 9 on the second test set of the first sub-task, and rank 4 on the test set of the second sub-task.

1 Introduction

Arabic is the official language of 22 countries, recognized as the 4th most used language on the Internet (Guellil et al., 2021). Arabic can be classified into three types (Guellil et al., 2021), mainly Classical Arabic (CA), Modern Standard Arabic (MSA), and Arabic Dialects (AD). Unlike both CA and MSA, Arabic Dialects lack a standardized representation and data that cover their complex taxonomy. Several initiatives were made to advance the research in this field. One of the most prominent work has been carried out through the Nuanced Arabic Dialect Identification (NADI) shared tasks. The first two NADI shared tasks (Abdul-Mageed et al., 2020, 2021b) comprised country-level and province-level dialect identification.

Many participants presented their systems to the NADI shared tasks. Most of the systems submitted rely on the Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019) models. For instance, Mansour et al. (2020) pretrained a multilingual BERT model on unlabeled Arabic tweets, then fine-tuned the model for the dialect classification task. Furthermore, Tahssin et al. (2020) fine-tuned the transformer-based Model for Arabic Language Understanding AraBERT (Antoun et al.) on an extended corpus constructed using a reverse translation of the given Arabic NADI dataset. Gaanoun and Benelallam (2020) employed Arabic-BERT (Safaya et al., 2020) alongside semi-supervised learning and ensembling methods in their system. El Mekki et al. (2020) introduced an ensemble that applies a weighted voting technique on two classifiers, the first based on TF-IDF with word and character n-grams and the second based on AraBERT. El Mekki et al. (2021) proposed a multi-task model that leverages MARBERT's contextualized word embedding (Abdul-Mageed et al., 2021a) with two task-specific attention layers, aggregated to predict both the province and the country of a given Arabic tweet.

The NADI 2022 shared task (Abdul-Mageed et al., 2022) provides two sub-tasks, mainly country-level dialect identification and sentiment analysis. Inspired by the previous submissions, we fine-tune AraBERT for each sub-task. The system for the first sub-task is a multi-task model that performs dialect identification by predicting the region, area, and country of the tweet. The system for the second sub-task is a distilled model from an ensemble of K models that were trained using K-fold cross-validation for sentiment classification.

This paper is structured as follows: Section 2 describes the data used. Section 3 gives an overview of fine-tuning BERT models. Section 4 presents the systems submitted to Subtasks 1 and 2 respectively. We show the results on the NADI Subtasks 1 and 2 and discuss them in Sections 5 and 6. We conclude with Section 7.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 485 - 490
December 8, 2022 ©2022 Association for Computational Linguistics

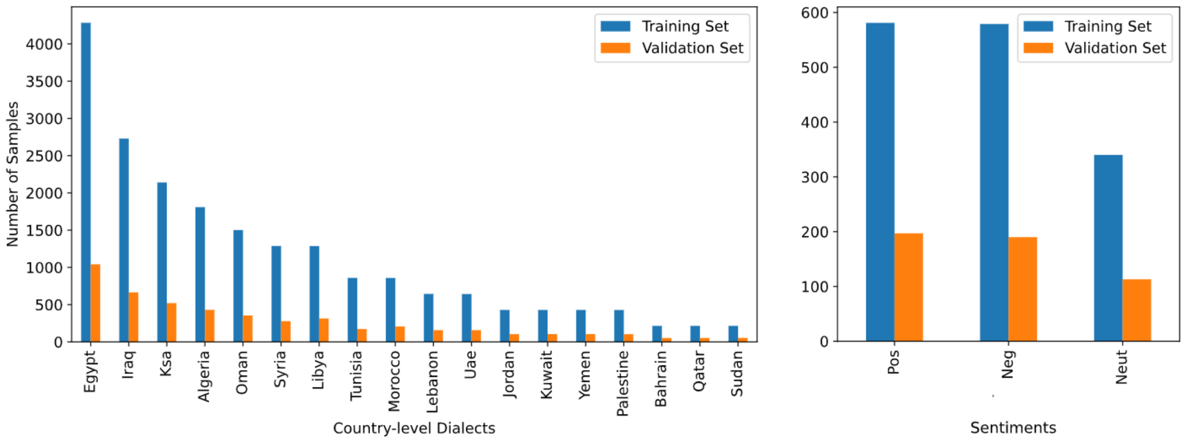


Figure 1: Label distribution in the training and validation sets of Subtask 1 and Subtask 2 respectively.

2 Data

2.1 Dataset Description

The systems were developed using the training and validation data provided by the task organizers. The training set for Subtask 1 consists of around 20,398 tweets with 18 different labels representing 18 country dialects, while the development set consists of 4,871 labeled tweets. The system submitted to this sub-task is evaluated on two test sets; the first test set (TEST-A) consists of 4,758 tweets covering 18 country-level dialects, whereas the second test set (TEST-B) consists of 1,474 tweets covering k country-level dialects.

The training set for Subtask 2 consists of 1,500 tweets labeled as either positive, negative, or neutral, while the development/validation set consists of 500 labeled tweets. The system submitted to this sub-task is evaluated on a test set of 3,000 unlabelled tweets.

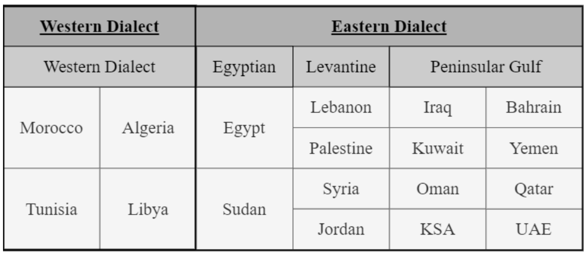
Figure 1 shows that the distribution of the tweets for the country-level classification sub-task is highly unbalanced. This would raise some issues in correctly predicting the minority classes (i.e., the dialects that have a small sample of tweets in the training set). Moreover, Figure 1 shows that the number of samples in the training set provided for the second sub-task is quite small. This raises the need to have a language model that can perform the task given the small training set. This motivates the use of transfer learning and pre-trained language models for this sub-task.

2.2 Dataset Pre-processing

We apply the same pre-processing techniques for both Subtask 1 and 2. We first standardize the text by removing non-Arabic words, emojis, and URLs from the tweets. Then, we proceed by tokenizing the tweets using the AraBERT tokenizer.

2.3 Region and Area Inference

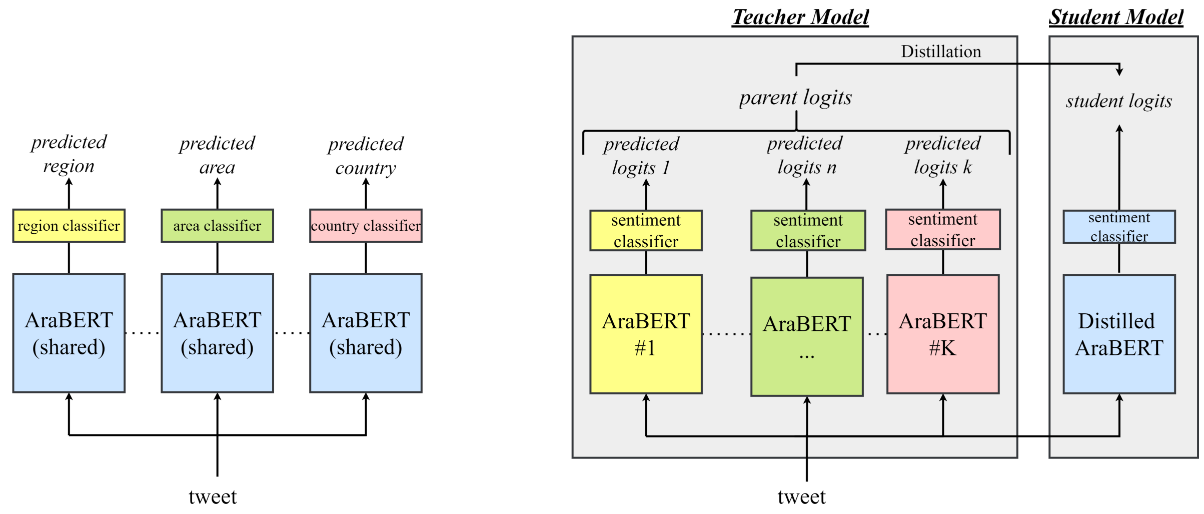
For the first sub-task, we infer two additional labels from the country-level label provided. We propose to classify the tweets into two regions (either Western or Eastern) and into four areas (Western, Egyptian, Levantine, or Peninsular gulf), as shown in Figure 2.



Western Dialect		Eastern Dialect			
Western Dialect		Egyptian	Levantine	Peninsular Gulf	
Morocco	Algeria	Egypt	Lebanon	Iraq	Bahrain
			Palestine	Kuwait	Yemen
Tunisia	Libya	Sudan	Syria	Oman	Qatar
			Jordan	KSA	UAE

Figure 2: Two additional labels were inferred from the country-level label for Subtask 1.

For instance, Morocco, Algeria, Tunisia, and Libya will belong to the Western region and to the Western area, while Egypt and Sudan will belong to the Eastern region and to the Egyptian area. We chose to add these additional labels to the task to encode some domain knowledge in the pre-trained language model.



(a) System for Subtask 1

(b) System for Subtask 2

Figure 3: Systems used for the NADI subtasks.

3 Fine-tuning BERT

As mentioned in the previous sections, the two subtasks fall under the category of text classification. An intuitive solution would be to fine-tune a pre-trained language model on each sub-task by adding an output layer to the encoder and training the parameters of the network to predict the classes for the subtask.

Fine-tuning is a form of Transfer Learning, as it tailors the knowledge encoded in the model to the downstream task. Therefore, it is crucial to find an appropriate model to fine-tune. After investigating multiple BERT variants, we choose to use an Arabic pre-trained language model called AraBERT (Antoun et al.). AraBERT is trained on a huge corpus of Arabic text from a collection of publicly available large-scale raw Arabic text. The specific model employed in both subtasks is the *bert-large-arabertv02-twitter*. It is based on AraBERTv0.2-large, and it is pre-trained using the Masked Language Modeling task on 60M Multi-Dialect Tweets.

However, fine-tuning a BERT variant might not be sufficient to reach the desired performance on the sub-tasks. Therefore, our contribution lies in employing multi-task learning for the first sub-task, and knowledge distillation from an ensemble of models for the second sub-task. All models have been trained on NVIDIA Tesla Volta V100.

4 Proposed Solutions

4.1 Subtask 1 - Multi-Task Learning

As mentioned in the previous section, a simple solution would be to fine-tune AraBERT to predict one dialect out of the 18 predefined dialects. We propose to encode more domain knowledge in AraBERT by training the model to predict the region and area of the tweet (as described in Figure 2). Learning these two labels jointly with the country-level dialect will help BERT acquire more knowledge for the country-level dialect identification task. To learn the region, area, and country-level dialect classes, we use multi-task learning. The Multi-Task model consists of a single shared AraBERT encoder. The pre-trained AraBERT model is fine-tuned using three task-specific classification heads (i.e., layers). Each classification head consists of a dropout layer of probability 0.1 followed by a linear layer that maps the CLS token embeddings of the AraBERT encoder to the number of predicted classes (2 classes for region, 4 for area, and 18 for country). We use the cross-entropy loss to compute the loss on the outcome of every classifier head. There are multiple strategies to combine the three losses. Since the losses assess different measures, we chose to fine-tune one loss at a time per batch. As seen in Figure 1, the dataset suffers from class imbalance. Therefore, we propose to randomly sample (with replacement) 500 sentences per country-level dialect. In other terms, the training set used for this model consists of 500

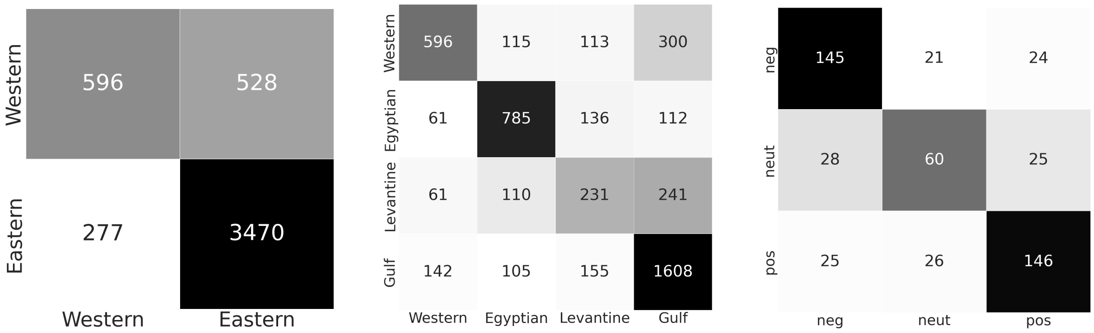


Figure 4: Confusion matrices for Region and Area of Subtask 1, and Sentiment of Subtask 2 on the dev set.

tweets for every label. This will guarantee that all classes participate in the training process equally. The model is trained using the Adam optimizer (Kingma and Ba, 2015), with a learning rate of 10^{-5}. After conducting multiple experiments, we chose to set the batch size to 64 and the number of epochs to 5. In this study, we report the results of the system that achieved the best score on the leaderboard.

4.2 Subtask 2 - Distilled Ensemble of K models

The proposed system relies on the same AraBERT model employed before. We propose to build an ensemble of K AraBERT models. To do so, we split the training set into K folds and we fine-tune an AraBERT model for each combination of (K-1) folds. Then, the output of this ensemble of K models (i.e., logits) is constructed by computing the average of the logits from all the K models. Using an ensemble is more robust and prevents overfitting since each model from that ensemble is exposed to a different subset of the training set. Furthermore, ensembles are known to usually achieve better performance compared to a single model. Afterward, we distill the knowledge from the ensemble teacher model to a single AraBERT student model by optimizing the following loss:

$$\begin{aligned} Loss &= (1-\alpha) \times CE(score, target) \\ &+ \alpha \times MSE(student\_logits, teacher\_logits) \end{aligned}$$

CE stands for cross-entropy loss, while MSE stands for mean squared error loss. We set α to 0.95 and K to 10. The model is trained with a learning rate of 5×10^{-6} and a batch size of 32 for 6 epochs. It should be noted that the hyperparameters reported are the ones that resulted in the best performance on the validation set.

5 Results

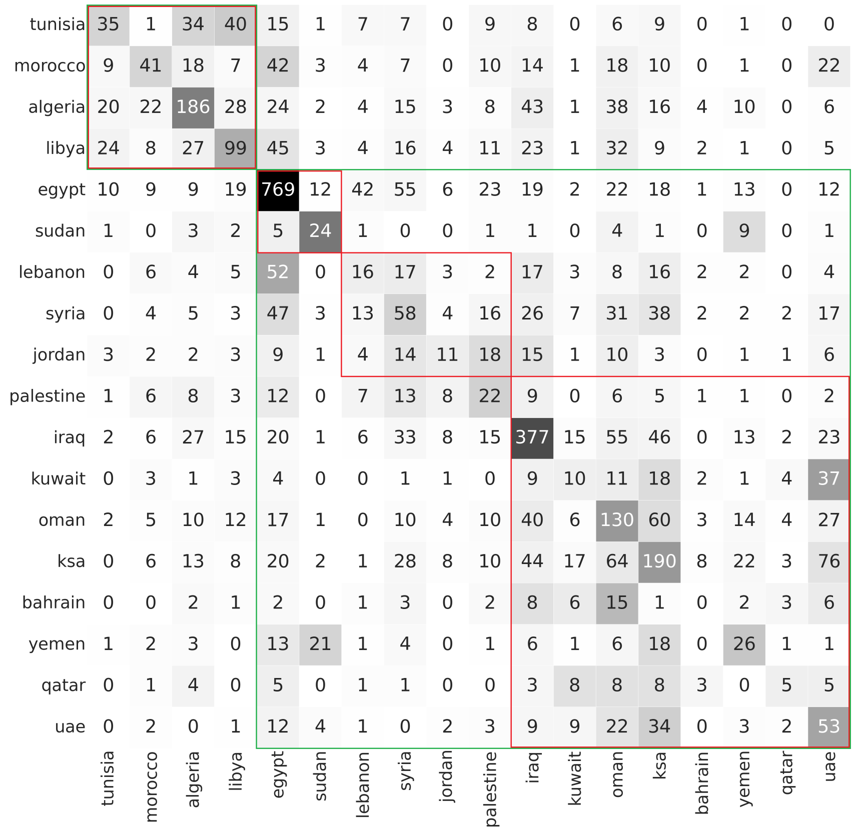
We evaluate our systems on the validation set provided by the organizers. Table 1 presents the Macro-Averaged Precision, Recall, and F1 Score computed over the development sets and reported on the test set by the organizers for each sub-task. The first sub-task was evaluated on two test sets TEST-A and TEST-B: TEST-A covers 18 country-level dialects, while TEST-B covers k country-level dialects, where k was kept unknown. The second sub-task was evaluated by computing the metrics over the positive and negative labels only, on one test set of 3000 tweets. The official metric used is the Macro-Averaged F1-score. We report the confusion matrices of both systems on the development sets in Figures 4 and 5.

Table 1: Results of the systems on Subtasks 1 and 2.

Sub -task	**Eval Set**	**Label**	**Macro Precision**	**Macro Recall**	**Macro F1 Score**
1	DEV	Region	77.53	72.81	74.64
		Area	61.80	60.17	60.65
		Country	28.50	28.01	27.57
	TEST-A	Country	36.77	31.77	32.63
	TEST-B	Country	19.51	15.90	15.61
2	DEV	Sentiment (Pos, Neg)	68.06	67.84	67.93
	TEST	Sentiment (Pos, Neg)	66.08	65.87	73.40

6 Discussion

As we can notice, the simple task of predicting whether the dialect is Western or Eastern is challenging by itself. This clearly confirms that the task of dialect identification is not an easy task. Furthermore, we notice that the model has trouble distinguishing between the Levantine dialect and the Peninsular Gulf dialect. This is expected as these dialects are the most similar among all four families (area).



	tunisia	morocco	algeria	libya	egypt	sudan	lebanon	syria	jordan	palestine	iraq	kuwait	oman	ksa	bahrain	yemen	qatar	uae
tunisia	35	1	34	40	15	1	7	7	0	9	8	0	6	9	0	1	0	0
morocco	9	41	18	7	42	3	4	7	0	10	14	1	18	10	0	1	0	22
algeria	20	22	186	28	24	2	4	15	3	8	43	1	38	16	4	10	0	6
libya	24	8	27	99	45	3	4	16	4	11	23	1	32	9	2	1	0	5
egypt	10	9	9	19	769	12	42	55	6	23	19	2	22	18	1	13	0	12
sudan	1	0	3	2	5	24	1	0	0	1	1	0	4	1	0	9	0	1
lebanon	0	6	4	5	52	0	16	17	3	2	17	3	8	16	2	2	0	4
syria	0	4	5	3	47	3	13	58	4	16	26	7	31	38	2	2	2	17
jordan	3	2	2	3	9	1	4	14	11	18	15	1	10	3	0	1	1	6
palestine	1	6	8	3	12	0	7	13	8	22	9	0	6	5	1	1	0	2
iraq	2	6	27	15	20	1	6	33	8	15	377	15	55	46	0	13	2	23
kuwait	0	3	1	3	4	0	0	1	1	0	9	10	11	18	2	1	4	37
oman	2	5	10	12	17	1	0	10	4	10	40	6	130	60	3	14	4	27
ksa	0	6	13	8	20	2	1	28	8	10	44	17	64	190	8	22	3	76
bahrain	0	0	2	1	2	0	1	3	0	2	8	6	15	1	0	2	3	6
yemen	1	2	3	0	13	21	1	4	0	1	6	1	6	18	0	26	1	1
qatar	0	1	4	0	5	0	1	1	0	0	3	8	8	8	3	0	5	5
uae	0	2	0	1	12	4	1	0	2	3	9	9	22	34	0	3	2	53

Figure 5: Confusion matrix for the country-level labels of Subtask 1.

Moreover, we notice that the confusion between dialects within the same area is higher compared to dialects from different areas (highlighted in Figure 5 by the clusters of values in red and green). This is expected as the training process injected knowledge that helps the model distinguish between the dialect classes (i.e., regions and areas). Therefore, a more fine-grained region-level and area-level classification should result in an improvement to the country-level dialect identification task.

We can also note the discrepancy in the performance of the model between TEST-A and TEST-B. In fact, TEST-A tests the model's performance on all the dialects, while TEST-B tests the performance on a subset of k dialects. TEST-B does not reflect the model's performance on all dialects, as the model might be tested on country-level dialects that are more difficult to predict.

As for Subtask 2, we can see that the Macro-Averaged F1 Score reported on the test set is higher than the score reported on the development set. This implies that distilling an ensemble of K models trained on different partitions of the training set helped the model generalize well on unseen data.

7 Conclusion

In this paper, we introduced two AraBERT-based systems to tackle dialect and sentiment classification. We conclude by confirming that dealing with Arabic dialect data is quite challenging. In future work, we propose to vary the training approach for every individual model in the ensemble, by changing the sequence length used, or even the training batch size per model. We also propose to build an ensemble of K multi-task models for Subtask 1.

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 Natu-*

ral Language Processing Workshop (WANLP 2020), Barcelona, Spain.

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 (WANLP 2021)*.

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. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of BERT and n-GRAM features for nuanced Arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274, Barcelona, Spain (Online). Association for Computational Linguistics.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. BERT-based multi-task model for country and province level MSA and dialectal Arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kamel Gaanoun and Imade Benelallam. 2020. Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 275–281, Barcelona, Spain (Online). Association for Computational Linguistics.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Moataz Mansour, Moustafa Tohamy, Zeyad Ezzat, and Marwan Torki. 2020. Arabic dialect identification using BERT fine-tuning. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 308–312, Barcelona, Spain (Online). Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Rawan Tahssin, Youssef Kishk, and Marwan Torki. 2020. Identifying nuanced dialect for Arabic tweets with deep learning and reverse translation corpus extension system. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 288–294, Barcelona, Spain (Online). Association for Computational Linguistics.