Beyond Borders: Uncovering Dialectal Arabic Overlaps through Multi-Label Identification

Anonymous ACL submission

Abstract

This paper explores Multi-Label Arabic Dialect Identification, addressing the limitations of single-label classification, which fails to capture the natural overlap between dialects. We use pseudo-labeling to generate multi-label training data and fine-tune BERT-based models to improve dialect classification. Our approach achieves state-of-the-art performance, surpassing previous methods by 7% in macro F1 score. These results show that allowing multiple dialect labels provides a more accurate representation of real-world language use. However, distinguishing similar dialects remains a challenge, emphasizing the need for better annotation techniques.

1 Introduction

Arabic is spoken by over 420 million people across more than 28 nations, is a highly diverse language encompassing Modern Standard Arabic (MSA) and a wide range of regional and national dialects. While MSA serves as the standardized form used in formal settings such as education, media, and official communication, Arabic dialects dominate informal interactions, including social media, text messaging, and everyday conversation. These dialects exhibit significant linguistic variation influenced by geography, culture, and history, making dialect identification a cornerstone challenge in Arabic Natural Language Processing.

Traditionally, Arabic Dialect Identification (ADI) has been framed as a single-label classification task, where a given text is associated with one dialect from a predefined set. However, this approach faces several challenges, as highlighted by (Keleg and Magdy, 2023). Short sentences often lack sufficient cues to indicate a single dialect, and MSA overlaps with all dialects, further complicating the task. Moreover, Arabic dialects exhibit significant diversity across regional (e.g., Levant, Gulf, and Maghreb), country (over 20 Arab nations), and city levels (more than 100 micro-dialects). Distinguishing between dialects at finer levels remains particularly challenging due to these overlaps, with ADI models consistently struggling to achieve robust performance, as evidenced by low macro-F1 scores.

Single-label classification models are limited in capturing the linguistic realities of Arabic dialects, where sentences often belong to multiple dialects. Overlapping expressions, shared vocabulary, and code-switching with languages such as French or English further complicate the task. These challenges have led to a paradigm shift in the field toward Multi-Label Dialect Identification (MLDID), where sentences can be associated with multiple dialects. Table 1 illustrates such overlaps, highlighting the difficulties of single-label classification. Furthermore, single-label datasets often introduce bias, as annotators tend to favor their native dialects, and shared linguistic features further blur distinctions.

Dialects	**Sentence**
Iraq, Jordan, Lebanon, Libya, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen	وين المحطة؟ (Where is the station?)
Iraq, Morocco, Qatar	شنو رقم الرحلة؟ (What is the flight/trip number?)
Lebanon, Syria, Jordan	شو عم تعمل؟ (What are you doing?)
Saudi Arabia, UAE, Qatar, Bahrain, Kuwait	شلونك؟ (How are you?)
All Arabic dialects	الله أكبر (Allah is great)

Table 1: Examples illustrating dialect overlap in Arabic.

The NADI 2024 shared task exemplifies this shift by introducing the MLDID subtask, which focuses on multi-label classification of country-level

Arabic dialects (Abdul-Mageed et al., 2024). To address these challenges, we introduce **B2BERT**, a transformer-based model specifically designed for **MLDID**. Our contributions can be summarized as follows:

- We propose a **synthetic multi-label dataset** by pseudo-labeling existing single-label dialect datasets, enabling models to learn overlapping dialectal features.
- We develop **B2BERT**, which is fine-tuned on this dataset and leverages **curriculum-based training approach** to improve classification accuracy, mitigating the impact of dataset imbalances.
- We demonstrate that **B2BERT achieves state-of-the-art (SOTA) performance** in the new paradigm of Arabic dialect identification, surpassing the top-performing models in the **NADI 2024 shared task leaderboard** with a **macro F1-score of 59.63%**.

Our work establishes a **new benchmark** for Arabic dialect identification and highlights the potential of **multi-label classification frameworks** in capturing the linguistic diversity of Arabic dialects.

2 Related Work

The task of Arabic Dialect Identification has emerged as a critical challenge in the field of Natural Language Processing (NLP) due to the vast linguistic diversity of Arabic dialects (Zaidan and Callison-Burch, 2014). This diversity, while rich in cultural and historical significance, poses significant obstacles for NLP applications, particularly with the widespread use of dialectal Arabic in digital communication, social media, and various online platforms. The ability to accurately identify and process these dialects is essential for enhancing communication technologies, developing more inclusive AI systems, and improving language-based applications like translation and sentiment analysis.

Recent years have witnessed a leap in research efforts aimed at tackling ADI. Early approaches primarily modeled ADI as a single-label classification problem (Abdul-Mageed et al., 2020; Zirikly et al., 2016; Bouamor et al., 2019), where each text sample was associated with a single dialect label. One of the main challenges of single-label ADI models is their inability to handle linguistic overlap across dialects. For instance, short sentences or common expressions may be valid in multiple dialects but are restricted to a single label in conventional datasets. Studies, such as Keleg and Magdy (2023), have demonstrated that approximately 66% of predictions classified as errors by single-label models are, in fact, valid in the predicted dialect. This reveals a critical evaluation bottleneck, as traditional metrics fail to account for the multi-dialectal nature of Arabic. Moreover, Althobaiti (2020) emphasized the biases introduced during manual annotation, where annotators often over-identify their native dialect, further skewing dataset validity.

Efforts like the Multi-Dialectal Parallel Corpus of Arabic (MPCA) (Bouamor et al., 2014) and the MADAR corpus (Bouamor et al., 2018) provided significant resources for dialect identification but were constrained by their reliance on single-label paradigms. These datasets, often constructed through manual or automated annotation techniques, fail to capture the intricate multi-label dynamics of dialectal texts.

Recognizing these limitations, the research community has started to advocate for reframing ADI as a multi-label classification task. NADI 2023's first subtask focused mainly on ADI in a single-label manner (Abdul-Mageed et al., 2023). Although this approach has led to some promising results from the teams participating in NADI Subtask 1 2023 (Elkaref et al., 2023); (Abdel-Salam, 2023); (Almarwani and Aloufi, 2023), these results were limited due to the reasons listed above. There have also been several efforts to develop parallel corpus datasets to efficiently capture the characteristics of each dialect. However, these datasets were parallelly translated from other languages rather than being naturally occurring, like tweets.

In NADI 2024(Abdul-Mageed et al., 2024), the focus shifted to MLDID. The emergence of this task posed serious challenges due to the nature of the dataset: each tweet in the dataset had a single label, but the objective was to generate multiple labels as the model's output.

Significant work was carried out by several teams participating in this task, particularly the work by (Karoui et al., 2024), who achieved the highest results by leveraging multi-label architectures, such as transformer-based models adapted for multi-output predictions. These models significantly improved evaluation fairness and perfor-

mance metrics, particularly in handling sentences with high dialectal ambiguity. Our goal is to improve these outcomes using new approaches.

3 Dataset

We used NADI 2020 (Abdul-Mageed et al., 2020), NADI 2021 (Abdul-Mageed et al., 2021b), and NADI 2023 (Abdul-Mageed et al., 2023), all of which provided tweets with single-label annotations. Furthermore, we incorporated the development set from NADI 2024's first subtask into our data pool. All these datasets were shared with us by the organizers.

The NADI 2023 dataset which is particularly notable for its balanced distribution, includes equal representation from 18 Arabic dialects from countries such as Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, UAE, and Yemen. Each dialect class in this dataset consists of 1000 tweets. In contrast, the datasets from NADI 2020 and NADI 2021 were found to be unbalanced, with dialects like Bahraini and Qatari being underrepresented compared to the more frequently encountered Egyptian and Iraqi dialects. Furthermore, classification in these datasets was influenced by the location from which posts were made, introducing a significant margin of error. The combined distribution for the three datasets is shown in Figure 1.

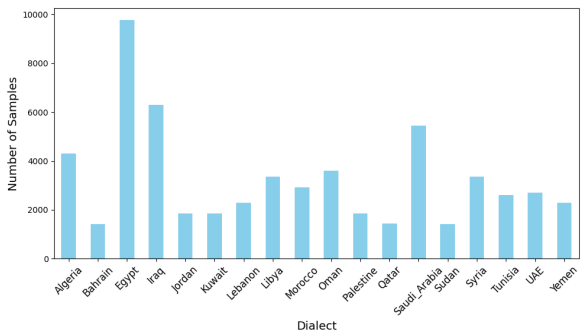

Figure 1: Number of samples in each dialect after combining the three datasets

To construct a more diverse and representative dataset we selected 31,760 samples from these three datasets. Since NADI 2023 dataset was balanced, it was used in full. Evenly sampled records were selected from NADI 2020 and NADI 2021 to enhance dataset consistency and avoid excessive dominance of particular dialects.

For evaluation, we utilized the official NADI 2024 test set which consists of 1,000 samples covering 14 Arabic-speaking countries This dataset provides a more reliable benchmark for assessing model performance on real-world dialect identification. The final dataset splits are provided in Table 2.

Splits	Sentences	Classes
Train	31,760	18
Validation	120	8
Test	1,000	14

Table 2: Final Dataset Splits.

4 Methodology

4.1 Baseline

We use the same baseline as that employed in the NADI 2024 (Abdul-Mageed et al., 2024) shared task. Specifically, a softmax-based model was fine-tuned for single-label dialect classification and adapted for the multi-label setting. The model outputs softmax probabilities for each dialect. It predicts the most probable labels until their cumulative probability exceeds 90%, allowing multiple dialects to be assigned to a sentence.

4.2 Proposed Method

The proposed approach consists of two main steps. First, a multi-label dataset is created from the original mono-label dataset by applying pseudo-labeling (Lee et al., 2013), allowing sentences with overlapping dialectal features to be assigned multiple labels. Pseudo-labeling is a semi-supervised learning technique where a model is initially trained on labeled data, then used to generate artificial (pseudo) labels for unlabeled data, treating them as ground truth for further training. This approach effectively expands the labeled dataset while leveraging the model's existing knowledge. Second, a multi-label classification model is fine-tuned on the generated dataset to predict all relevant dialects for a given sentence.

4.2.1 Dataset Creation

The provided NADI dataset is originally mono-labeled across 18 Arabic dialects.

Three distinct approaches were explored for converting the mono-label dataset into a multi-label dataset. **The first approach employed a pipeline of logistic regression classifiers**, where

18 independent classifiers were trained to determine whether a sentence belonged to a specific dialect. Tweets labeled with the target dialect, along with MSA sentences, were treated as positive samples, while negative samples were evenly selected from the remaining dialects to balance training data. This approach served as an initial attempt at multi-label classification, leveraging the efficiency of logistic regression.

The second approach improved upon this by fine-tuning MARBERT (Abdul-Mageed et al., 2021a) binary classifiers, a classifier for each dialect. MARBERT is an Arabic specific BERT based model. MARBERT's ability to capture nuanced dialectal differences made it a strong candidate for binary classification at a more fine-grained level. Each binary classifier was fine-tuned using the same dataset setup as the logistic regression approach, ensuring a fair comparison between models.

The third approach utilized GPT-4 for pseudo-labeling, leveraging large language models (LLMs) to generate high-quality multi-label annotations. Carefully crafted prompts were designed to guide GPT-4 in detecting dialectal features in tweets and assigning appropriate dialect labels. This approach allowed us to explore how LLMs perform in dialect identification, particularly in cases where dialect boundaries are unclear.

An analysis of the pseudo-labeled dataset revealed a significant imbalance, with most samples containing only a single active label. Multi-label samples, particularly those with 16 or more active labels, were underrepresented. To mitigate this, instances with 16 and 17 active labels were merged with the 18-label category (representing MSA).

4.2.2 Multi-Label Classification

The initial experiments focused on a pipeline of binary classifiers trained on the original single-label dataset. Two approaches were explored: one using 18 fine-tuned MARBERT (Abdul-Mageed et al., 2021a) classifiers, each predicting a specific dialect, and another using logistic regression classifiers trained independently for each dialect. While these models provided a basic framework for dialect prediction, they were inherently limited by the constraints of single-label classification, failing to capture the overlapping nature of dialects in many sentences.

To address this limitation, we fine-tuned multi-label classification models using the pseudo-labeled multi-label dataset. MARBERT and CAMeLBERT (Inoue et al., 2021) were selected for this task due to their pretraining on dialectal Arabic datasets, which made them well-suited for dialect identification. However, initial results were sub-optimal, revealing that the pseudo-labeled dataset was highly imbalanced, with most samples containing only a small number of active labels (ranging from 1 to 3 dialects). This imbalance skewed the model's learning process, making it less effective at identifying multi-dialect samples.

To mitigate this issue, an undersampling strategy was applied, focusing on reducing the dominance of samples with few active labels. The goal was to ensure that the model was trained on a more balanced distribution of label combinations. While this adjustment improved overall performance, a new challenge emerged: the model struggled to accurately classify sentences with a higher number of active labels (sentences that belong to many dialects). This suggested that after undersampling, predicting samples with a larger number of active labels became more difficult.

This observation motivated the adoption of curriculum-based training (Soviany et al., 2022), a technique inspired by human learning processes, where models are trained on progressively structured examples rather than being exposed to all complexities at once. Given that after undersampling, the model struggled with higher active labels, we hypothesized that introducing them gradually, rather than all at once, would help mitigate this issue. Initially, the model was trained exclusively on samples containing a single active label to establish a strong foundation. In subsequent epochs, samples with two active labels were introduced alongside a proportional number of single-label samples. This gradual inclusion continued across epochs, introducing samples with higher numbers of active labels at each step while maintaining balanced representation across categories. By the final epoch, the model had been exposed to the full range of label complexities, enabling it to generalize effectively and handle complex multi-label scenarios. This approach ensured balanced learning without the drastic data reduction caused by undersampling.

By structuring the training in this manner, the model was encouraged to progressively adjust to more complex label distributions, preventing it from being overwhelmed by high-active-label samples too early. This provided a systematic alterna-

tive to standard undersampling, preventing excessive information loss while ensuring that the model learned to recognize complex multi-label patterns more effectively.

5 Experiments and Evaluation

To ensure a fair comparison among the experimented systems, we adopted the same hyper-parameters used by the top-performing team in NADI2024 (Karoui et al., 2024) shared task. These settings, listed in Table 3, were carefully selected to optimize model performance while maintaining consistent evaluation metrics across different configurations.

The experiments were conducted using a single NVIDIA RTX 6000 GPU with 24GB of memory. The training was performed with a batch size of 11 and ran for 10 epochs. Each experiment took approximately 27 minutes to complete. The models fine-tuned include MARBERT, which has approximately 163M parameters, and CAMeLBERT, which consists of about 110M parameters. Using these standardized hyperparameters ensured a fair and direct comparison with the best-performing system in the shared task, while also maintaining computational efficiency and consistency across our experiments.

Hyper-parameter	Value
Learning Rate	1e-05
Optimizer	AdamW
Train Batch Size	11
Evaluation Batch Size	11
Number of Training Epochs	10
Dropout Rate	0.3

Table 3: Fine-Tuning Hyper-parameters.

5.1 Dataset Preprocessing

The preprocessing stage focused on cleaning and standardizing text from the NADI 2020, 2021, and 2023 datasets. Specific cleaning was applied to the 2021 dataset to remove @ symbol before the USER and https before the URL placeholder tags.

We removed punctuation, emojis, and diacritics to reduce noise, with URLs and mentions replaced by placeholders to retain context while guaranteeing anonymization for the training data. Character normalization was applied by unifying Alef variants, for example: converting 'ى' to 'ي' and 'ة' to 'ه', and reducing repeated letters (e.g., 'ااا') to a single occurrence to avoid inflating word counts. Stopwords were eliminated using an Arabic and English stopwords list to focus on meaningful text.

We also addressed mixed-language text, introducing spaces between Arabic and English characters to prevent parsing issues caused by language switching then removed all english text. After cleaning each dataset individually, we concatenated them into a single dataset to be used in our specific task. This preprocessing ensured the data was consistent, normalized, and ready for model training and evaluation.

We utilized Python libraries such as NLTK (Bird and Loper, 2004), Camel Tools (Obeid et al., 2020), and PyArabic (Zerrouki, 2023) to perform text cleaning and normalization for Arabic text.

6 Results

To assess the performance of our models on the multi-label classification task, we utilized macro F1-score, precision, recall, and accuracy as evaluation metrics. These metrics provide a comprehensive evaluation of each model's effectiveness, particularly in handling overlapping dialects and distinguishing between similar ones. The experiments highlighted the strengths and limitations of various models and training strategies under different dataset configurations.

We present the performance of our MLDID pipeline across the most significant experiments, showcasing the impact of dataset creation strategies such as undersampling and curriculum-based training. Each experiment was evaluated using the scoring methodology of the NADI shared task to ensure consistency and comparability. As benchmarks, we included the ELYADATA model (Karoui et al., 2024) and NADI2024 shared task baseline.

All reported results in the following sections are based on the development set. A final evaluation on the test set is presented separately.

6.1 Experiments on Binary and Multi-Label Classifiers

We first evaluate traditional single-label dialect classification using independent binary classifiers for each dialect. The results from table **??** indicate that MARBERT outperforms logistic regression significantly, highlighting the advantage of transformer-based models in dialect identification.

Models were also trained on pseudo-labeled datasets generated using three different ap-

	F_1	P	R
Binary Classifiers Pipeline using Single-label Data			
Logistic Regression (18)	0.4018	0.4389	0.4027
MARBERT (18)	0.5841	0.5559	0.6752
Logistic Regression Pseudo Labels			
CAMeLBERT	0.5238	0.4680	0.6437
MARBERT	0.5755	0.5576	0.6370
Logistic Regression Pseudo Labels (Undersampled)			
CAMeLBERT	0.5477	0.5125	0.6260
MARBERT	0.5730	0.5267	0.6709
MARBERT Classifiers Pseudo Labels			
CAMeLBERT	0.5808	0.4882	0.8052
MARBERT	0.5527	0.4780	0.7263
MARBERT Classifiers Pseudo Labels (Undersampled)			
CAMeLBERT	0.5884	0.4767	0.8573
MARBERT	0.5729	0.4948	0.7772
GPT4 Pseudo-Labels			
CAMeLBERT	0.4534	0.6233	0.3908
MARBERT	0.4600	0.7588	0.3688
GPT4 Pseudo-Labels (Undersampled)			
CAMeLBERT	0.6411	0.7090	0.6357
MARBERT	0.5597	0.7141	0.5153
Curriculum-Based Training - GPT Pseudo-Labels			
CAMeLBERT(B2BERT)	**0.6549**	**0.6966**	**0.6552**
MARBERT	0.6532	0.6896	0.6543

Table 4: Performance Comparison of Models Based on Macro-Average Scores

proaches: LR-based pseudo-labeling, MARBERT-based pseudo-labeling, and GPT-4 pseudo-labeling. For each dataset, two multi-label classification models, MARBERT and CAMeLBERT, were fine-tuned. The dataset configurations included the whole dataset and the undersampled dataset. Additionally, for the GPT-4 pseudo-labeled dataset, curriculum-based training was applied as a third configuration.

The results for models trained on the LR-based pseudo-labeled dataset show that CAMeLBERT exhibited a small improvement when trained on the undersampled dataset, increasing from 0.5238 to 0.5477 in macro F1-score. However, MARBERT showed a slight drop, with its score decreasing from 0.5755 to 0.5730. These results indicate that while undersampling helped CAMeLBERT slightly, it did not provide a consistent benefit for MARBERT, suggesting that balancing strategies alone may not fully address the challenges of multi-label dialect identification.

A similar pattern is observed in models trained on the MARBERT pseudo-labeled dataset. CAMeLBERT improved slightly with undersampling, increasing from 0.5808 to 0.5884, while MARBERT showed a more noticeable gain, rising from 0.5527 to 0.5729. This suggests that pseudo-labeling quality plays a bigger role in model performance than dataset balancing alone, as both models performed better than their LR pseudo-labeled counterparts.

Among all dataset variations, the GPT-4 pseudo-labeled dataset produced the highest macro F1 scores across different models and dataset configurations. Unlike the other two datasets, undersampling led to significant improvements, particularly for CAMeLBERT, which increased from 0.4534 to 0.6411. However, MARBERT showed a less significant improvement, rising from 0.4600 to 0.5597, indicating that while balancing the dataset helped, the model still struggled with certain dialectal variations.

Applying curriculum-based training led to further performance improvements, with MARBERT achieving its highest macro F1-score of 0.6532 and CAMeLBERT reaching 0.6549. This underscores curriculum-based training as the most effective strategy for enhancing generalization, particularly when applied to high-quality pseudo-labeled data.

These findings highlight the importance of high-quality pseudo-labeling, where GPT-4-generated labels consistently outperformed both LR and MARBERT-based pseudo-labeling. Furthermore, the success of curriculum-based training suggests that models benefit from a gradual increase in label complexity, particularly when applied to datasets with rich dialectal variations. The results also show that while undersampling improved performance for certain models, it was not a universally effective solution, reinforcing the need for structured training approaches.

6.1.1 Final Evaluation on the Test Set

For completeness, the final performance of selected models is evaluated on the test set. Table 5 summarizes the macro F1-scores, precision, and recall for B2BERT (CAMeLBERT + GPT-4 Pseudo-labeled data + Curriculum-learning), ELYADATA (Karoui et al., 2024), and the baseline model.

Model	Macro F1	P	R
NADI2024 Baseline	0.4698	0.648	0.3986
ELYADATA	0.5240	0.5015	0.5687
B2BERT	**0.5963**	**0.5818**	**0.6976**

Table 5: Final Performance on the Test Set

The test set evaluation provides an objective comparison between our best-performing model and existing benchmarks.

6.2 Discussion

To assess the performance of our models on the multi-label classification task, we analyzed both quantitative metrics and qualitative examples. This approach highlights key challenges and areas for improvement. One notable challenge is the difficulty in distinguishing between dialects in the Maghreb region. For instance, the sentence 'عيش خويا، مريقل نتقابلو مبعد كان هكا' (Live, brother, everything is fine, we'll meet later if so) is a pure Tunisian dialect. However, the model incorrectly predicted Algeria and Morocco alongside Tunisia. This confusion suggests that the model struggles to capture subtle linguistic differences between closely related dialects.

A similar issue arises with Sudanese Arabic. For the sentence 'كل زول ليه الزول بتاعه' (Everyone has their own person), which is clearly Sudanese, but the model mistakenly included Egypt as a predicted label. This misclassification may be attributed to the annotation methodology employed during the dataset creation. This methodology focuses on location metadata.

On a more positive note, the model demonstrated strong performance in identifying MSA. For instance, the sentence 'الله أكبر' (Allah is the greatest) was correctly classified with all relevant labels activated, showcasing the model's robustness in handling MSA. This indicates that the implemented curriculum learning approach successfully strengthened the model's ability to generalize to less ambiguous cases while gradually introducing complexity during training.

Overall, the results demonstrated that while the model shows strong potential in handling multi-label classification tasks for Arabic dialects, it still faces challenges in differentiating closely related varieties. The model also encounters difficulties in accurately classifying sentences that are characteristic of a single dialect, often incorrectly assigning them to multiple dialects. This tendency to over-generalize suggests that the model may struggle to discern the nuanced linguistic features that distinguish each dialect. These difficulties emphasize the need to enhance the model's ability to capture subtle linguistic and contextual cues specific to each dialect, even within closely related groups. Furthermore, addressing inconsistencies in the dataset, such as noise introduced by geographic overlaps or metadata-driven annotations, could significantly improve the model's accuracy and generalizability.

7 Conclusion and Future Work

In this study, we introduced B2BERT, a model designed to tackle the challenge of Multi-Label Arabic Dialect Identification by recognizing the natural overlap between dialects. By using GPT-4-based pseudo-labeling and curriculum-based training, B2BERT effectively learns from imbalanced dialect distributions, achieving a macro F1-score of 59.63% and outperforming all previous approaches, including the top-performing model in the NADI 2024 shared task.

As a next step, we aim to refine annotation techniques to ensure cleaner labels, explore data augmentation to strengthen generalization, and expand the model's coverage to include a broader range of dialects. Additionally, incorporating self-training and semi-supervised learning could allow the model to make better use of unlabeled data, further enhancing its accuracy.

Ultimately, this work provides a strong foundation for improving ADI, bringing us closer to language technologies that better reflect the richness and diversity of the Arabic-speaking world.

Limitations

The model has several limitations, which are discussed in this section. Firstly, the NADI dataset was annotated based on geographic regions, which may introduce noise as some gold labels might be inaccurately assigned. To mitigate this issue, we propose engaging dialect experts to review and, if necessary, correct the dataset annotations. Additionally, implementing a multi-annotator system could ensure that each sample is reviewed multiple times, increasing the inter-annotator agreement and enhancing the overall quality of dataset curation. Secondly, the conversion of the dataset from single-label to multi-label, our primary contribution, was not performed manually. This introduces potential errors in the multi-label dataset, which could neg-

atively impact the model's performance in certain scenarios where dialects overlap or share similarities. The model may struggle to distinguish closely related dialects; for example, if a sentence is purely Tunisian, the model might incorrectly classify it as Tunisian, Moroccan, and Algerian.

8 Ethics and Broader Impact

Human Subject Considerations. All annotators provided informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time.

Transparency and Reproducibility. To promote open research, we release our code to the public.

References

Reem Abdel-Salam. 2023. rematchka at NADI 2023 shared task: Parameter efficient tuning for dialect identification and dialect machine translation. In *Proceedings of ArabicNLP 2023*, pages 652–657, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. Arbert marbert: Deep bidirectional transformers for arabic.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. 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.

Nada Almarwani and Samah Aloufi. 2023. SANA at NADI 2023 shared task: Ensemble of layer-wise BERT-based models for dialectal Arabic identification. In *Proceedings of ArabicNLP 2023*, pages 625–630, Singapore (Hybrid). Association for Computational Linguistics.

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

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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, Florence, Italy. Association for Computational Linguistics.

Mohab Elkaref, Movina Moses, Shinnosuke Tanaka, James Barry, and Geeth Mel. 2023. NLPeople at NADI 2023 shared task: Arabic dialect identification with augmented context and multi-stage tuning. In *Proceedings of ArabicNLP 2023*, pages 642–646, Singapore (Hybrid). Association for Computational Linguistics.

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.

Amira Karoui, Farah Gharbi, Rami Kammoun, Imen Laouirine, and Fethi Bougares. 2024. ELYADATA at NADI 2024 shared task: Arabic dialect identification with similarity-induced mono-to-multi label transformation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 758–763, Bangkok, Thailand. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey.

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

Taha Zerrouki. 2023. Pyarabic: A python package for arabic text. *Journal of Open Source Software*, 8(84):4886.

Ayah Zirikly, Bart Desmet, and Mona Diab. 2016. The GW/LT3 VarDial 2016 shared task system for dialects and similar languages detection. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 33–41, Osaka, Japan. The COLING 2016 Organizing Committee.