ABDUL: a new Approach to Build language models for Dialects Using formal Language corpora only

Yassine Toughrai[1,2], Kamel Smaïli[1,2], David Langois[1,2]
[1]Université de Lorraine
[2] Laboratoire lorrain de recherche en informatique et ses applications
{yassine.toughrai, smaili, david.langlois}@loria.fr

Abstract

Arabic dialects present major challenges for natural language processing (NLP) due to their diglossic nature, phonetic variability, and the scarcity of resources. To address this, we introduce a phoneme-like transcription approach that enables the training of robust language models for North African Dialects (NADs) using only formal language data, without the need for dialect-specific corpora. Our key insight is that Arabic dialects are highly phonetic, with NADs particularly influenced by European languages. This motivated us to develop a novel approach in which we convert Arabic script into a Latin-based representation, allowing our language model, ABDUL, to benefit from existing Latin-script corpora. Our method demonstrates strong performance in multi-label emotion classification and named entity recognition (NER) across various Arabic dialects. ABDUL achieves results comparable to or better than specialized and multilingual models such as DarijaBERT, DziriBERT, and mBERT. Notably, in the NER task, ABDUL outperforms mBERT by 5% in F1-score for Modern Standard Arabic (MSA), Moroccan, and Algerian Arabic, despite using a vocabulary four times smaller than mBERT.

1 Introduction

NADs, including Moroccan, Algerian, and Tunisian, introduce additional complexities. Influenced by Berber languages and colonial languages such as French and Spanish, these dialects display notable phonetic variability, including vowel inconsistency and the adoption of phonemes absent in MSA, such as /p/ and /v/ (Barkat-Defradas et al., 2003). In addition, their lexicons are enriched by extensive borrowing from French and Spanish and often incorporating them with phonetic modifications (Owens, 2013).

In this article, we introduce a phoneme-like transcription approach that bridges formal Arabic with dialectal varieties through linguistic normalization. Inspired by the Buckwalter (Buckwalter, 2002) transliteration system, our method simplifies and adapts transliteration by clustering phonetically similar sounds, improving alignment with dialectal phonetic patterns. To highlight consonants and long vowels (e.g., the "ā" in the word kitāb for "book" which is pronounced with an extended duration of the vowel /a/), this approach deliberately omits diacritization and even removes preexisting diacritics from the text, reducing phonetic variability (Al-Mozainy, 1981).

By transforming Arabic script into a standardized phoneme-like Latin representation, this preprocessing pipeline promotes cross-script and cross-dialect generalization, allowing for the development of robust NLP models trained solely on formal language data. In this article, we will focus exclusively on transliterating MSA to handle Arabic dialects, with the future goal of including French and code-switched text, given their significance in NADs.

2 Linguistic Justification

NADs are low-resource languages with no formal or standardized grammatical rules, relying mainly on direct phonetic transcription. Alongside MSA vocabulary, they feature extensive lexical borrowings from French, Spanish, Turkish, and Italian, reflecting the historical and colonial influence of these languages in the region. The lexical resemblance between Algerian Arabic (ALG) and MSA has been quantitatively analyzed using computational methods. Abukwaik et al. (Abu Kwaik et al., 2018) employed Latent Semantic Indexing (LSI) to assess lexical overlap between MSA and various Arabic dialects, reporting an LSI similarity score of 0.68 for Algerian Arabic. This score indicates a moderate lexical divergence, suggesting that while some vocabulary is shared, directly applying MSA-

trained models to Algerian Arabic could result in significant tokenization mismatches. Harrat et al. (Harrat et al., 2014) found that approximately 20% of Algerian dialectal words originate from Arabic, while 34% are derived from MSA. Studies estimate that loanwords make up around 30–40% of the vocabulary in these dialects, particularly in technical, educational, and governmental contexts (Owens, 2013; Barkat-Defradas et al., 2003). In (Harrat et al., 2016), the authors argue that significant variations between and MSA occur in vocalization, along with the omission or modification of certain letters, particularly the Hamza[1]. Despite the influence of foreign lexicons, NADs preserve core linguistic structures from MSA. However, in terms of pronunciation, Menacer et al. (Menacer et al., 2017) found that 46% of MSA-derived words in NADs exhibit phonetic variations compared to their standard MSA counterparts. Another key characteristic of NADs is their strong dependence on consonantal structures for lexical and semantic distinctions, as vowel patterns vary significantly across regions (Barkat-Defradas et al., 2003). Given these linguistic properties, ABDUL leverages stable consonantal structures, which serve as robust subword units for training NLP models, reducing variability caused by inconsistent usage of vowels.

3 Related Work

NADs are low-resource languages that lack formalized grammatical rules and primarily rely on phonetic transcription. In this work, we propose a novel paradigm for training language models for NADs using only formal language corpora, eliminating the need for dialect-specific datasets. To evaluate the effectiveness of our approach, we compare it against several key baselines in Arabic NLP, particularly those designed for Arabic dialects:

- **AraBERT** [2]: A pretrained BERT model for MSA (Antoun et al., 2020), serving as a foundational model for Arabic NLP. It is trained on a mix of MSA corpora and Arabic Wikipedia, capturing linguistic nuances in formal Arabic.
- **mBERT** [3]: A multilingual BERT model pretrained on 100+ languages (Devlin et al., 2019). While not specifically optimized for Arabic, it provides a multilingual perspective on cross-lingual transfer.
- **DarijaBERT** [4]: A BERT model fine-tuned for Moroccan Arabic (Darija) (Gaanoun et al., 2023), leveraging localized datasets to capture dialect-specific nuances.
- **TunBERT** [5]: A Tunisian Arabic BERT model (Messaoudi et al., 2021), highlighting the lexical and phonological idiosyncrasies of this dialect.
- **DziriBERT** [6]: A pretrained model for Algerian Arabic (Dziri) (Abdaoui et al., 2022), providing a benchmark for North African dialectal NLP.

Beyond these baselines, our approach is further inspired by the study "Consonant is All You Need" (Al-shaibani and Ahmad, 2023), which highlights the benefits of reducing reliance on vowels for more efficient NLP models. This work demonstrates how selectively omitting certain lexical features can lead to smaller vocabularies, lower computational complexity, and improved training efficiency. These insights align with our dediacritization and consonant-centric transcription strategy, reinforcing the scalability and effectiveness of our method.

Through a rigorous comparative analysis, we aim to underscore the advantages of our approach. By pretraining a language model from scratch on data processed via our pipeline, we establish a fair and consistent benchmark to demonstrate the benefits of phoneme-like transcription for Arabic dialectal NLP. Our work contributes to the broader goal of improving low-resource language modeling through linguistically informed methodologies.

4 Methodology

To effectively adapt formal Arabic resources for dialectal NLP, we develop a preprocessing pipeline that normalizes phonetic variability while preserving linguistically significant features. In the following, we outline the key steps in our phoneme-like transcription process.

4.1 Phoneme-like Transcription Pipeline

Our preprocessing pipeline transforms Arabic text into a phoneme-like Latin representation by:

[1]The Hamza is a letter in the Arabic alphabet representing the glottal stop

[2]https://huggingface.co/aubmindlab/bert-base-arabertv2

[3]https://huggingface.co/google-bert/bert-base-multilingual-uncased

[4]https://huggingface.co/SI2M-Lab/DarijaBERT

[5]https://huggingface.co/tunis-ai/TunBERT

[6]https://huggingface.co/alger-ia/DziriBERT

1. **Dediacritization**: We remove short vowels and diacritics, treating them as having a minimal impact to normalize phonetic variations and highlight consonant structures. This approach aligns with the principle that consonants encode the fundamental semantic meaning of words in Arabic (Watson, 2002).

2. **Retention of Long Vowels**: long vowels are preserved to capture essential phonetic cues while reducing ambiguity, reflecting their phonemic stability in Arabic dialects (Al-Mozainy, 1981).

3. **Simplified Transliteration**: Inspired by the Buckwalter transliteration system, our simplified Latin-script transcription ensures phonetic consistency across dialects. This improves tokenization efficiency and allows models trained with formal Arabic corpora to generalize better to dialects, particularly by unifying phonetically similar sounds under a shared representation.

4.2 Model Training

We pretrain a BERT model from scratch using the Arabic split of the OSCAR corpus (Ortiz Suárez et al., 2019), applying our preprocessing pipeline. We utilize a WordPiece tokenizer with a vocabulary size of 30,522 tokens. The model undergoes training for 9 epochs using the Adam optimizer, with a learning rate of 5e-5, a batch size of 64, and a maximum sequence length of 512. Training is conducted on a single NVIDIA A100, with a masked language modeling (MLM) probability of 0.15.

The choice of vocabulary size plays a crucial role in language model training, especially for morphologically rich languages like Arabic. To ensure fair comparison, we adopt the BERT architecture, aligning with benchmark models such as DarijaBERT, DziriBERT, TunBERT, AraBERT, and mBERT. The 30,522-token vocabulary was selected to match the lowest vocabulary size among these benchmarks (Table 1), allowing for an equitable evaluation of efficiency across different pretraining settings.

5 Datasets

In this section, we describe the datasets used for pretraining and benchmarking ABDUL, covering MSA and NADs. Our selection includes a large-scale corpus in MSA for pretraining and multiple dialect-specific datasets for downstream tasks, ensuring a comprehensive evaluation across emotion classification and named entity recognition (NER).

Table 1: Vocabulary size comparison between the ABDUL trained BERT model and the models it will be benchmarked against

Language	Model	Vocab Size
Moroccan	DarijaBERT	80,000
Algerian	DziriBERT	50,000
Tunisian	TunBERT	30,522
MSA	arabert	64,000
Multilingual	mBERT	119,547
MSA	ABDUL	30,522

5.1 Pretraining Dataset

We use the Arabic subset of the OSCAR corpus (Ortiz Suárez et al., 2019) for pretraining our model. This dataset contains approximately 8.7 million documents and 6.1 billion words, totaling around 84.2 GB of text. Derived from web sources such as news articles, blogs, and forums, OSCAR provides a diverse representation of MSA. Its scale and domain diversity make it well-suited for training transformer-based language models, ensuring broad linguistic coverage.

5.2 Emotion Classification Datasets

For text classification, we employ the SemEval 2025 [7] Task 11-A dataset, which focuses on emotion detection in Moroccan and Algerian Arabic. The dataset consists of approximately 900 labeled instances per dialect, annotated with four emotion categories: joy, anger, sadness, and fear. This dataset serves as a benchmark for evaluating emotion classification in different research works concerning NADs, which pose unique linguistic challenges due to their phonetic variations and lexical borrowings.

5.3 Named Entity Recognition (NER) Datasets

For NER evaluation, we utilize three datasets: WikiFANE(Alotaibi and Lee, 2014), DzNER (Dahou and Cheragui, 2023), and DarNER (Moussa and Mourhir, 2023), which cover different dialects and entity types, providing a comprehensive benchmark for dialectal Arabic NER.

[7] https://semeval.github.io/SemEval2025/

- **WikiFANE**: Covers MSA and NADs, providing a general-purpose dataset.

- **DzNER**: Focuses on Algerian Arabic, with a broader range of entity types.

- **DarNER**: Specializes in Moroccan Arabic and includes date entities in addition to standard entity categories.

Table 2 summarizes the dataset attributes.

6 Results

The performance of ABDUL is evaluated on two tasks: emotion classification and named entity recognition (NER), across multiple Arabic variants. The results demonstrate ABDUL's ability to generalize effectively across dialects while maintaining competitive performance against specialized models.

6.1 Emotion Classification Performance

Table 3 presents the emotion classification results for Algerian Arabic. ABDUL achieves a macro-F1 score of **0.5315**, ranking second behind DziriBERT (**0.5573**). It outperforms DarijaBERT (**0.5107**), the specialized Moroccan Arabic model, and significantly surpasses TunBERT (**0.2473**), which struggles in this dialect.

Table 3: Emotion classification results for Algerian Arabic.

Model	Precision	Recall	Macro-F1	Accuracy
DarijaBERT	0.6454	0.4289	0.5107	0.2747
DziriBERT	**0.6560**	0.4928	**0.5573**	**0.3186**
TunBERT	0.4220	0.2210	0.2473	0.2087
arabert	0.6369	0.4159	0.4964	0.2417
mBERT	0.5295	0.3434	0.4071	0.2197
ABDUL	0.6000	**0.5014**	0.5315	0.2088

Table 4 presents the emotion classification results for Moroccan Arabic. ABDUL achieves a macro-F1 score of **0.4519**, closely matching AraBERT (**0.4518**), a model trained on MSA. It outperforms DarijaBERT (**0.4648**) and significantly surpasses TunBERT (**0.1020**).

Table 4: Emotion classification results for Moroccan Arabic.

Model	Precision	Recall	Macro-F1	Accuracy
DarijaBERT	0.5399	**0.4122**	**0.4648**	**0.5280**
DziriBERT	0.5057	0.3589	0.4157	0.4410
TunBERT	0.1538	0.0797	0.1020	0.2981
arabert	**0.7039**	0.3775	0.4518	0.4907
mBERT	0.4109	0.2777	0.3254	0.3727
ABDUL	0.5266	0.4035	0.4519	0.4596

Table 5 presents the averaged classification results across dialects. ABDUL achieves an overall macro-F1 score of **0.4915**, outperforming DarijaBERT (**0.4878**) and DziriBERT (**0.4865**). This highlights ABDUL's ability to generalize across NADs despite being trained exclusively on MSA.

Table 5: Average emotion classification results across dialects.

Model	Precision	Recall	Macro-F1	Accuracy
DarijaBERT	0.5926	0.4205	0.4878	**0.4013**
DziriBERT	0.5808	0.4258	0.4865	0.3798
TunBERT	0.2879	0.1503	0.1746	0.2535
arabert	**0.6704**	0.3967	0.4741	0.3662
mBERT	0.4702	0.3106	0.3662	0.2962
ABDUL	0.5633	**0.4524**	**0.4915**	0.3342

These results suggest that ABDUL's phoneme-like transcription preprocessing effectively captures dialectal features while avoiding reliance on extensive dialect-specific data. Its particularly strong performance in Algerian Arabic underscores its suitability for handling underrepresented dialects in emotion classification.

6.2 Named Entity Recognition (NER) Performance

Table 6 presents the results for NER in MSA. ABDUL achieves an F1 score of **0.4646**, performing on par with mBERT (**0.4647**), the top-performing model. It surpasses arabert (**0.4427**), demonstrating its effectiveness in formal Arabic settings. The results assess ABDUL's ability to generalize across different Arabic variants and effectively capture named entities despite phonetic and lexical variability.

Table 2: NER datasets for Arabic and North African dialects.

Dataset	Language/Dialect	Entities	Size in tokens
WikiFANE	MSA and North African Dialects	102 different entities	490k
DzNER	Algerian Arabic (Darija)	PER, LOC, ORG, MISC	220k
DarNER	Moroccan Arabic (Darija)	PER, LOC, ORG, DATE	65,905

Table 6: NER performance on MSA.

Model	Precision	Recall	F1	Accuracy
DarijaBERT	0.4922	0.4112	0.4481	0.8927
DziriBERT	0.4825	0.3854	0.4285	0.8911
TunBERT	0.4416	0.0068	0.0134	0.8633
arabert	0.5016	0.3962	0.4427	0.8954
mBERT	0.5152	**0.4233**	**0.4647**	0.8977
ABDUL	**0.5180**	0.4212	0.4646	**0.8979**

For NER in Algerian Arabic (Table 7), ABDUL achieves the highest F1 score of **0.6828**, significantly outperforming DziriBERT (**0.5461**), which is specifically trained for this dialect.

Table 7: NER performance on Algerian Arabic.

Model	Precision	Recall	F1	Accuracy
DarijaBERT	0.5556	0.5615	0.5585	0.9384
DziriBERT	0.5361	0.5565	0.5461	0.9382
TunBERT	0.4286	0.0071	0.0140	0.9104
arabert	0.5104	0.5841	0.5448	0.9389
mBERT	0.4975	0.5727	0.5325	0.9343
ABDUL	**0.6601**	**0.7071**	**0.6828**	**0.9553**

For NER in Moroccan Arabic (Table 8), ABDUL attains an F1 score of **0.6557**, ranking just behind mBERT (**0.7192**), the best-performing model overall. However, it surpasses DarijaBERT (**0.6246**), that was designed especially for Morrocan. A qualitative analysis of the DarNER corpus revealed that many words were transcribed in a way that closely aligns with their Arabic root rather than reflecting phonetic pronunciation. This likely explains mBERT's and arabert's superior performance, as these models benefit from their extensive pretraining on MSA.

Table 8: NER performance on Moroccan Arabic.

Model	Precision	Recall	F1	Accuracy
DarijaBERT	0.6077	0.6424	0.6246	0.9272
DziriBERT	0.5875	0.5516	0.5690	0.9193
TunBERT	0.1928	0.0548	0.0853	0.8436
arabert	0.6491	0.6761	0.6623	0.9290
mBERT	**0.7140**	**0.7246**	**0.7192**	**0.9403**
ABDUL	0.6415	0.6706	0.6557	0.9346

Table 9 presents the averaged NER performance across MSA, Algerian, and Moroccan Arabic. ABDUL achieves an overall F1 score of **0.6010**, outperforming both DarijaBERT (**0.5437**) and DziriBERT (**0.5145**). This demonstrates ABDUL's ability to generalize effectively across dialects while maintaining strong performance in both formal and informal Arabic varieties.

Table 9: Average NER performance across Arabic variants.

Model	Precision	Recall	F1	Accuracy
DarijaBERT	0.5518	0.5384	0.5437	0.9194
DziriBERT	0.5354	0.4978	0.5145	0.9162
TunBERT	0.3543	0.0229	0.0376	0.8724
arabert	0.5564	0.5521	0.5500	0.9211
mBERT	0.5756	0.5735	0.5721	0.9241
ABDUL	**0.6065**	**0.5996**	**0.6010**	**0.9293**

7 Conclusion

ABDUL consistently matches or exceeds the performance of specialized models for certain dialects in tasks such as emotion classification and named entity recognition (NER), despite being trained exclusively on MSA. It notably outperforms DarijaBERT and DziriBERT in several scenarios, showcasing its strong adaptability to NADs. By utilizing a phoneme-like transcription approach, ABDUL effectively bridges the gap between formal and dialectal Arabic, improving tokenization efficiency and enhancing generalization across dialects with shared linguistic features. Its ability to compete with dialect-specific models while relying solely on widely available, high-quality MSA data underscores its scalability and potential for low-resource Arabic NLP.

8 Limitations and Future Work

While ABDUL demonstrates strong performance in dialectal NLP tasks, several limitations remain. Currently, our approach does not support Latin-script Arabizi dialects, which are widely used in informal settings. Expanding ABDUL to handle Arabizi is a key part of our future work. Additionally, we plan to investigate how vocabulary size

impacts performance, as well as how different formal languages used in pretraining (e.g., French, Spanish, and English) influence the model's ability to generalize across dialects.

Overall, the results are low for state-of-the-art models, including ABDUL. The task will be to test other architectures to improve the results and not settle for the current ones.

Finally, we aim to expand ABDUL's applicability to a broader set of NLP tasks, including machine translation and text generation, to further assess its scalability and effectiveness in diverse linguistic contexts. As a long-term objective, we seek to build the first large language model (LLM) for Arabic dialects, leveraging the high availability and quality of formal languages data to address the low-resource status of Arabic dialects and advance the field of dialectal Arabic NLP.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2022. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13.

Hamza Al-Mozainy. 1981. *Vowel Alternations in a Bedouin Hijazi Arabic Dialect: Abstractness and Stress*. Ph.D. thesis, University of Texas at Austin.

Maged S. Al-shaibani and Irfan Ahmad. 2023. Consonant is all you need: a compact representation of english text for efficient NLP. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Fahd Alotaibi and Mark Lee. 2014. A hybrid approach to features representation for fine-grained Arabic named entity recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 984–995, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Melissa Barkat-Defradas, Jalal Al-Tamimi, and Thami Benkirane. 2003. Phonetic variation in production and perception of speech: a comparative study of two Arabic dialects. In *Proceedings of the 15th International Congress of Phonetic Sciences ((ICPhS), Barcelona 3-9 August 2003*, pages pp. 857–860, Barcelona, Spain.

Tim Buckwalter. 2002. Arabic transliteration. *URL http://www. qamus. org/transliteration. htm*.

Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023. Dzner: A large algerian named entity recognition dataset. *Natural Language Processing Journal*, 3:100005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, pages 4171–4186.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect.

Salima Harrat, Karima Meftouh, Mourad Abbas, Walid-Khaled Hidouci, and Kamel Smaïli. 2016. An Algerian dialect: Study and Resources. *International journal of advanced computer science and applications (IJACSA)*, 7(3):384–396.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building Resources for Algerian Arabic Dialects. In *15th Annual Conference of the International Communication Association Interspeech*, Singapour, Singapore. ISCA.

Mohamed Amine Menacer, Odile Mella, Dominique Fohr, Denis Jouvet, David Langlois, and Kamel Smaïli. 2017. Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. In *ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics*, pages 1–8, Dubai, United Arab Emirates.

Abir Messaoudi, Ahmed Cheikhrouhou, Hatem Haddad, Nourchene Ferchichi, Moez BenHajhmida, Abir Korched, Malek Naski, Faten Ghriss, and Amine Kerkeni. 2021. Tunbert: Pretrained contextualized text representation for tunisian dialect.

Hanane Nour Moussa and Asmaa Mourhir. 2023. Darnercorp: An annotated named entity recognition dataset in the moroccan dialect. *Data in Brief*, 48:109234.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Jonathan Owens. 2013. *Arabic as a Minority Language*.

Janet Watson. 2002. *Phonology and morphology of Arabic (the phonology of the world's languages)*.