Modeling North African Dialects from Standard Languages

Yassine Toughrai[1,2] **Kamel Smaïli**[1,2] **David Langlois**[1,2]
[1]Université de Lorraine
[2]Laboratoire Lorrain de Recherche en Informatique et ses Applications
{yassine.toughrai, smaili, david.langlois}@loria.fr

Abstract

Processing North African Arabic dialects presents significant challenges due to high lexical variability, frequent code-switching with French, and the use of both Arabic and Latin scripts. We address this with a phoneme-based normalization strategy (Toughrai et al., 2025) that maps Arabic and French text into a simplified representation (Arabic rendered in Latin script), reflecting native reading patterns. Using this method, we pretrain BERT-based models on normalized Modern Standard Arabic and French only and evaluate them on Named Entity Recognition (NER) and text classification. Experiments show that normalized standard-language corpora yield competitive performance on North African dialect tasks; in zero-shot NER, Ar_20k surpasses dialect-pretrained baselines. Normalization improves vocabulary alignment, indicating that normalized standard corpora can suffice for developing dialect-supportive language models in low-resource contexts.

1 Introduction

Arabic dialects are underrepresented in natural language processing (NLP), particularly in North African varieties such as Algerian, Moroccan, and Tunisian Arabic. These dialects are characterized by rich linguistic variation and frequent code-switching with French (Hamed et al., 2025), yet they lack sufficient labeled or unlabeled corpora to support robust language modeling. This scarcity hinders both tool development and the creation of reliable annotated datasets for downstream tasks.

In this work, we investigate whether strong representations for North African Arabic dialects can be learned using only standard (non-dialectal) data—namely Modern Standard Arabic (MSA) and French. We adopt a phoneme-oriented normalization that reduces surface-level divergence between dialects and MSA (e.g., vowel masking). By aligning lexical and orthographic variation in this way, we induce subword units with greater overlap between dialectal and standard tokens, enabling more consistent tokenization across varieties.

We pretrain a suite of BERT-style (Devlin et al., 2019) models using only MSA and French corpora. Our models differ in two key dimensions: (1) vocabulary size (20k, 30k, and 40k), and (2) pretraining data composition (Arabic only, Arabic + French, and Arabic + French with synthetic code-switched text). These controlled ablations allow us to assess how vocabulary granularity and multilingual exposure affect downstream generalization.

To evaluate the effectiveness of these models, we fine-tune them on dialectal Named Entity Recognition (NER) task and sentiment polarity classification task using publicly available datasets. The results show that several pretrained variants we developed, outperform strong baselines, including dialect-specific and multilingual models, despite having no access to dialectal data during pretraining.

Additionally, we perform a detailed out-of-vocabulary (OOV) analysis across datasets, demonstrating that even the smallest vocabulary (20k) achieves near-complete coverage, and suggesting that carefully normalized standard language corpora can yield high subword coverage for dialectal data, enabling effective downstream adaptation without dialectal pretraining.

Our work also contributes an evaluation framework that includes underused resources for Algerian and Moroccan Arabic, helping guide future benchmarking of North African dialect models. These findings point toward a promising direction for modeling Maghrebi Arabic dialects using standard Arabic resources alone, a setting underexplored in current research.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 375–383

2 Related Work

Recent work on Arabic NLP has prioritized the development of dialect-specific models, particularly for North African and Gulf varieties. Examples include DziriBERT (Abdaoui et al., 2021) for Algerian Arabic and TunBERT (Haddad et al., 2022) for Tunisian Arabic, both trained on large social media corpora. While effective, these models depend on dialectal pretraining resources that remain scarce, noisy, or fragmented for many dialects.

General-purpose MSA models such as AraBERT v2 (Antoun et al., 2020) and ARBERT (Abdul-Mageed et al., 2021) offer broader coverage but often struggle with dialectal input due to lexical and orthographic mismatch. Nonetheless, MSA-trained models have been shown to perform surprisingly well on dialectal tasks—especially when trained on undiacritized data (Abdul-Mageed et al., 2021; Antoun et al., 2020).

To reduce surface variation between MSA and dialects, prior work has explored character-level modeling and phonological normalization (Meftouh et al., 2015). Studies on diacritics restoration (Harrat et al., 2013; Mubarak et al., 2019) have further highlighted the differences between standard and dialectal Arabic and the benefits of simplification at the orthographic level.

This paper extends that line of work by introducing a surface harmonization technique, such as long-vowel abstraction (Toughrai et al., 2025) that unifies dialectal and standard forms at the token level. These techniques are applied not just as preprocessing but are used during pretraining, enabling the model to learn dialect-compatible representations while being exposed only to MSA (and French) language corpora.

Several recent studies have adapted MSA-pretrained models to dialects via light supervision. CAMeLBERT-DA (Inoue et al., 2021), for example, introduces adapter layers to specialize an MSA-pretrained model for individual dialects. It improves performance on dialectal NER and POS tagging tasks using lightweight fine-tuning. Similarly, Khalifa et al. (Khalifa et al., 2021) explore self-training for zero and few-shot dialectal adaptation, showing notable improvements on multi-dialect NER and POS.

However, these approaches still rely on access to dialectal data for adaptation. In contrast, our work adopts a zero-dialectal pretraining setting: we investigate whether models trained exclusively on surface-harmonized MSA and French corpora can generalize to dialectal tasks such as NER and polarity classification. This reflects a realistic low-resource scenario, where dialectal corpora are unavailable during pretraining and fine-tuning.

Despite the popularity of adaptation-based methods, few studies have directly compared dialectal pretraining with MSA-only pretraining for dialectal tasks. Most evaluations have focused on transfer learning without modifying the training corpus (El Mekki et al., 2021; Abdul-Mageed et al., 2021). Our study addresses this gap by showing that corpus-level surface harmonization enables robust dialectal transfer even without exposure to dialect data.

Finally, subword vocabulary size plays a crucial role in balancing coverage and generalization. Prior works on model compression and efficiency (e.g., DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2020)), as well as Algerian-specific modeling (Laggoun et al., 2025), suggest that smaller models can retain competitive performance through careful vocabulary and architecture choices. Our study complements this by comparing 20k, 30k, and 40k vocabulary sizes and showing that even the smallest configurations maintain strong coverage and downstream performance, especially when paired with surface-harmonized inputs.

Overall, our work introduces a scalable and linguistically motivated approach to dialectal modeling. By leveraging surface harmonization and pretraining on MSA and French only, we demonstrate that strong performance on dialectal tasks can be achieved without access to dialectal data during pretraining—offering a viable strategy for modeling under-resourced Arabic varieties.

3 Model Pretraining

We adopt a BERT-style encoder architecture rather than an autoregressive decoder model (e.g., GPT), as our primary objective is to learn robust, transferable representations for downstream dialectal tasks such as NER and polarity classification. Indeed, BERT's bidirectional context encoding is particularly effective in morphologically rich and syntactically flexible languages such as Arabic, where dialectal cues are often context-sensitive. This architecture enables token-level understanding over both left and right contexts, which is critical for fine-grained classification tasks on noisy, code-switched, or informal text.

To reduce dialectal variation and promote lexical alignment between Modern Standard Arabic (MSA) and dialects, we introduce a *phoneme-like normalization* strategy that maps Arabic text into a simplified, long-vowel-focused representation. Inspired by how Arabic readers naturally develop fluency without diacritics, we strip all short vowels and diacritics and merge phonetically similar letters, then transliterate as is to Latin script. For example, Arabic letters like t, T and v (written here in Buckwalter format) are all mapped to t, and long vowels such as A, w, and y (in Buckwalter) are retained. Weakly pronounced or orthographically unstable characters such as hamza (’), taa marbuta (p), and hamza-on-ya (}) are replaced with a generic placeholder (e.g., #). We give in Table 1 examples of the results of this normalization.

For French, we apply a comparable transformation by removing all vowels and retaining consonants, punctuation, and word boundaries. This creates a consonant-driven representation more structurally aligned with Arabic and facilitates subword vocabulary sharing, especially in code-switched settings.

All models were pretrained using a masked language modeling (MLM) objective following the BERT base architecture. pretraining was conducted for 10 epochs over approximately 128GB of text using three different GPU configurations, based on availability. All models used a maximum sequence length of 512 tokens and were optimized using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of 0.01. A linear learning rate scheduler was used with 10,000 warmup steps, followed by decay from a peak learning rate of 1e-4.

Buckwalter	Normalized	Arabic Script
AlHayApu	al7ya#	الحياة
manATiqu	mnatq	مناطق
Alba$ariỹapu	alb%ry#	البشرية

Table 1: Examples of phoneme-like normalization applied to Arabic text using transliteration.

We construct our pretraining corpus from three sources. First, we use the entire Arabic subset of the OSCAR 22.01 corpus (Abadji et al., 2022), comprising over 8.7 million documents and roughly 6.1 billion words. Second, we include 1 million French documents from the same source to reflect the prevalence of French borrowings in North African dialects. Third, we synthesize Arabic–French code-switched text using the Arabic–French portion of the OPUS OpenSubtitles dataset. Each Arabic sentence is aligned with its French translation, and we randomly select approximately 25% of Arabic content words (nouns, adjectives, named entities) to be translated into French using the Helsinki-NLP/opus-mt-ar-fr model on HuggingFace. We do not translate function words or morphologically complex verbs in order to preserve grammatical structure. This yields diverse and fluent code-switched sequences that reflect switching frequencies typical of informal Maghrebi discourse.

Tokenization is performed using the WordPiece algorithm, with subword vocabularies of size 20k, 30k, or 40k depending on the model variant. Vocabularies are trained jointly on the preprocessed Arabic and French data. All models are trained from scratch with no exposure to downstream dialectal datasets or evaluation labels during pretraining.

4 Evaluation

To assess the quality of the learned representations, we evaluate on two complementary classification tasks: Named Entity Recognition (NER) and sentiment polarity classification. NER probes token-level semantic and morphological information, while polarity classification targets sentence-level semantic and pragmatic understanding. For downstream fine-tuning, we train each baseline and proposed model for up to 20 epochs and select the single best checkpoint by macro-F1 on a held-out validation split comprising 10% of the training data; the test set is evaluated exactly once on this selected checkpoint. Models, preprocessing scripts, and full hyperparameters will be available HuggingFace[1].

4.1 Nomenclature

Throughout the paper, we refer to multiple model variants based on vocabulary size and pretraining setting. Results are reported for three vocabulary sizes: 20k, 30k, and 40k. Each variant is identified using the following notation:

- **{Ar}_{Xk}**: Models pretrained on Arabic-only (MSA normalized) data with a vocabulary size of *X* thousand tokens (e.g., Ar_20k).
- **{Ar+Fr}_{Xk}**: Models pretrained on a mix of Arabic (MSA) and French (normalized)

[1] https://huggingface.co/collections/YassineToughrai/abdul-pretrained-models-68cd78d6936fb6e90d7283fd

Model	Training Data Size	Language/Dialect	Vocab Size	Source Type
DarijaBERT	~100M tokens	Moroccan Arabic	80k	Tweets, YouTube, Stories
DziriBERT	~20M tokens	Algerian Arabic	50k	Tweets
TunBERT	~500k sentences	Tunisian Arabic	48k	Common Crawl
AraBERT v2	~1.5B words	Modern Standard Arabic	64k	Web (incl. OSCAR)
MARBERT	~128M tweets	Dialectal Arabic (multi-region)	64k	Twitter
CAMeLBERT-DA	Adapter tuning on MADAR	Multiple Dialects	64k	MSA+MADAR
mBERT	Wikipedia (100+ langs)	Multilingual (incl. Arabic)	110k	Wikipedia

Table 2: Overview of baseline models used for evaluation.

data with a vocabulary size of X thousand tokens (e.g., Ar+Fr_30k).

- **{Ar+Fr+CS}**: A fine-tuned Ar+Fr_40k on (normalized) synthetic code-switched text, using a vocabulary of 40k tokens.

This notation is used consistently across tables and figures for clarity and compactness.

4.2 Evaluation Setup

For the NER evaluation, we use three datasets: **DzNER** (Dahou and Cheragui, 2023), **DarNER** (Moussa and Mourhir, 2023), and **WikiFANE** (Alotaibi and Lee, 2013). **DzNER** contains over 21,000 Algerian social media sentences (approximately 220k tokens) collected from Facebook posts and YouTube comments, annotated with three entity types: PER, ORG, and LOC. **DarNER** is a Moroccan dialect dataset comprising 65,947 tokens extracted from Moroccan Arabic Wikipedia, annotated with four entity types: PER, ORG, LOC, and MISC. **WikiFANE** is a fine-grained NER dataset based on MSA Wikipedia articles, consisting of roughly 500,000 tokens labeled with 50 fine-grained entity classes. All NER models are fine-tuned for 20 epochs using a learning rate of 5×10^{-5}.

For the polarity classification evaluation, we use the **TwiFil** dataset, a collection of 9,000 Algerian Arabic tweets labeled for sentiment polarity. The authors of the corpus gathered the tweets between 2015 and 2019 and manually annotated them into three classes: positive, negative, and neutral. All polarity classification models are fine-tuned for 10 epochs using a learning rate of 2×10^{-5}.

We compare the performance of the pretrained models against a suite of strong Arabic and dialectal baselines that do not use normalization of training data. Table 2 summarizes the key properties of these models, including their training data sizes, dialectal focus, vocabulary sizes, and source types.

Obviously, in the experiments, the test data was normalized when using our models, and not normalized when using standard baseline models.

4.3 Vocabulary Coverage Analysis

To assess lexical coverage, we analyze the number and proportion of unknown tokens ([UNK]) produced during tokenization. This evaluation is conducted on the **test splits** of all datasets used in downstream evaluation: DarNER, DzNER, WikiFANE, and TwiFil.

On DzNER, all tokenizers, regardless of vocabulary size or language composition, produce exactly one unknown token out of more than 207,000 tokens. We observe the same pattern on WikiFANE (0 unknowns) and DarNER (131 unknowns), with no variation across tokenizers. These results suggest that even relatively small (20k) subword vocabularies trained solely on (normalized) MSA or MSA+French are sufficient to achieve near-complete lexical coverage of both standard language and dialectal text when combined with phoneme-like normalization.

Tokenizer	Total	UNKs	Rate (%)
Ar_20k	40,316	12	0.03
Ar_30k	38,696	12	0.03
Ar_40k	37,853	12	0.03
Ar+Fr_20k	40,603	14	0.03
Ar+Fr_30k	39,056	14	0.04
Ar+Fr_40k	38,075	14	0.04

Table 3: UNK Token Statistics on TwiFil

Similarly, Table 3 shows results for TwiFil, which consists of informal and noisy Algerian tweets. Here, unknown token rates are still very low (0.03–0.04%) but show minor variation across tokenizers. This may indicate that user-generated dialectal content presents slightly more challenges for tokenization, although the overall coverage remains high in absolute terms.

Tokenizers	Ar_20k	Ar_30k	Ar_40k	Ar+Fr_20k	Ar+Fr_30k	Ar+Fr_40k
Ar_20k	100.00	100.00	100.00	89.63	98.63	98.94
Ar_30k	65.53	100.00	100.00	59.36	87.78	98.74
Ar_40k	50.00	76.31	100.00	45.48	67.40	90.11
Ar+Fr_20k	89.63	90.59	90.97	100.00	100.00	100.00
Ar+Fr_30k	65.76	89.31	89.86	66.67	100.00	100.00
Ar+Fr_40k	48.83	74.37	88.95	49.36	74.03	100.00

Table 4: Percentage of subword vocabulary in each tokenizer (rows) that overlaps with another tokenizer (columns). Values are relative to the row tokenizer.

Similar trends are observed on DarNER and WikiFANE. These consistent results across datasets further support the robustness of phoneme-normalized tokenization in bridging standard and dialectal variation.

The limited effect of vocabulary size on unknown token rates can be partially explained by the overlap between vocabularies. As shown in Table 4, most of the additional subwords introduced in larger vocabularies are already present in the 20k base vocabulary. For instance, 98.63% of subwords in Ar_20k are also found in Ar+Fr_30k, and even Ar+Fr_40k retains a 90.11% overlap with Ar_40k. This high degree of redundancy likely contributes to the consistent tokenization behavior observed across vocabulary sizes.

In summary, even a 20k subword vocabulary trained on (normalized) standard Arabic and French yields high lexical coverage across all datasets. The consistently low OOV rates and minimal UNK variation suggest that our phoneme-like normalization strategy helps unify dialectal and standard language surface forms into shared subword units. While we do not directly measure alignment, the high vocabulary overlap and robust downstream performance indicate that normalization promotes structurally compatible tokenizations across varieties.

4.4 Evaluation Results

We evaluate the performance of our models and baseline models on two core classification tasks: sentiment polarity classification (semantic sentence level) and named entity recognition (token-level). We compare against a set of strong baseline models, including AraBERT v2 (Antoun et al., 2020)[2], which is partially pretrained on the Arabic portion of OSCAR 22.01, as well as DziriBERT (Abdaoui et al., 2021), DarijaBERT (Gaanoun et al., 2023), TunBERT (Haddad et al., 2022), MARBERT (Abdul-Mageed et al., 2021), CAMeLBERT-DA (Inoue et al., 2021) (adapted on MADAR (Bouamor et al., 2019)), and the multilingual BERT (mBERT)[3].

4.4.1 Polarity Classification

Model	Accuracy	F1 Score
DarijaBERT	69.64	68.96
DziriBERT	**73.68**	71.42
TunBERT	59.92	57.39
AraBERT v2	73.28	71.94
CAMeLBERT-DA	72.47	71.90
MARBERT	71.26	70.42
mBERT	68.42	67.67
Ar_20k	70.04	69.59
Ar_30k	69.64	67.29
Ar_40k	71.26	70.50
Ar+Fr_20k	72.47	70.00
Ar+Fr_30k	72.06	69.55
Ar+Fr_40k	72.87	**71.96**
Ar+Fr+CS	71.26	70.08

Table 5: Polarity Classification Results on TwiFil

We evaluate polarity classification on the TwiFil (Moudjari et al., 2020) dataset, which includes 9,000 Algerian Arabic tweets annotated into three sentiment classes: positive, negative, and neutral. All models are fine-tuned for 10 epochs with a learning rate of 2×10^{-5}. Table 5 reports accuracy and F1 scores.

As shown in Table 5, several pretrained models achieve strong results on the TwiFil dataset. Notably, **DziriBERT CAMeLBERT-DA** and **AraBERT v2** perform competitively, with F1 scores above 71, underscoring the effectiveness of dialect-specific and well-established MSA mod-

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

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

els on sentiment classification. However, our **Ar+Fr_40k** model surpasses all baselines in F1, suggesting that phoneme-informed pretraining on standard Arabic and French can yield robust generalization, even without access to dialectal corpora. Interestingly, we observe minimal variation across vocabulary sizes (20k, 30k, 40k), consistent with our OOV analysis showing negligible differences in unseen token rates. Furthermore, while the **Ar+Fr+CS** variant achieves solid results, it does not outperform the **Ar+Fr** models, indicating that explicit fine-tuning on synthetic code-switched data provides only modest additional benefit. This may suggest that some degree of cross-lingual alignment is already captured during pretraining, though further targeted analysis is needed to confirm this.

4.4.2 Named Entity Recognition

We further evaluate at a semantic but token-level the performance of our pretrained models using Named Entity Recognition (NER) on three datasets: WikiFANE, DzNER, and DarNER. All models are fine-tuned for 20 epochs with a learning rate of 5×10^{-5}. We report precision, recall, accuracy, and macro F1 scores.

Model	Acc	F1
DarijaBERT	89.58	44.63
DziriBERT	89.44	44.15
TunBERT	86.06	01.88
AraBERT v2	89.49	46.21
CAMeLBERT	89.73	47.83
MARBERT	**90.23**	47.66
mBERT	89.66	47.02
Ar_20k	89.74	46.57
Ar_30k	89.86	46.77
Ar_40k	90.00	46.87
Ar+Fr_20k	90.04	47.56
Ar+Fr_30k	89.95	47.12
Ar+Fr_40k	89.98	**47.92**
Ar+Fr+CS	89.92	47.72

Table 6: NER Performance on WikiFANE Dataset

Despite the overall strong results, performance on the WikiFANE dataset (Table 6) is lower than on DzNER (Table 7) and DarNER (Table 8). This is likely due to WikiFANE's substantially larger label space, with over 50 fine-grained entity types. The increased complexity of this classification task introduces greater potential for label confusion and sparsity across categories, making it more challenging for models to generalize effectively. In contrast, DzNER and DarNER contain fewer and coarser-grained entity classes, which reduces the prediction space and allows the models to perform more robustly.

Model	Acc	F1
DarijaBERT	93.89	55.77
DziriBERT	94.05	58.04
TunBERT	90.98	04.86
AraBERT v2	95.31	65.75
CAMeLBERT	92.47	62.98
MARBERT	93.58	67.23
mBERT	94.34	61.11
Ar_20k	95.93	71.75
Ar_30k	95.90	71.90
Ar_40k	96.02	72.13
Ar+Fr_20k	**96.25**	**74.60**
Ar+Fr_30k	95.95	72.38
Ar+Fr_40k	95.92	73.25
Ar+Fr+CS	95.92	72.48

Table 7: NER Performance on DzNER Dataset

Model	Acc	F1
DarijaBERT	93.24	65.21
DziriBERT	92.37	60.76
TunBERT	83.80	10.67
AraBERT v2	93.27	67.41
CAMeLBERT	92.84	53.66
MARBERT	**94.69**	61.90
mBERT	94.37	**72.83**
Ar_20k	94.01	70.83
Ar_30k	93.76	68.68
Ar_40k	93.87	70.28
Ar+Fr_20k	94.40	70.80
Ar+Fr_30k	94.44	71.39
Ar+Fr_40k	94.29	71.14
Ar+Fr+CS	94.06	70.67

Table 8: NER Performance on DarNER Dataset

Ablations (vocabulary and code-switching): The vocabulary-size ablation and OOV coverage analysis indicate that most lexical benefits are captured at 20k; larger vocabularies provide limited additional value in downstream NER. Fine-tuning on synthetic code-switched text (Ar+Fr+CS) yields modest changes but does not surpass the strongest Ar+Fr variants. While promising, these findings remain preliminary, and further controlled studies are needed to disentangle the individual effects of

Model	NER	All Tasks	Rank ↓
Ar+Fr_40k	64.10	**66.07**	**1.75**
Ar+Fr_20k	**64.32**	65.74	4.75
Ar+Fr+CS	63.62	65.24	4.75
Ar+Fr_30k	63.63	65.11	5.50
Ar_40k	63.09	64.94	6.25
Ar_20k	63.05	64.68	7.50
Ar_30k	62.45	63.66	9.00
AraBERT	59.79	62.83	7.75
mBERT	60.32	62.16	7.75
MARBERT	58.93	61.80	7.25
CAMeLBERT-DA	54.82	59.09	7.00
DarijaBERT	55.20	58.64	11.50
DziriBERT	54.32	58.59	10.25
TunBERT	5.80	18.70	14.00

Table 9: Average F1 and ranks across the supervised tasks (per-dataset F1 ranks; 1 = best)

vocabulary size, training data, and normalization strategies.

Cross-dataset summary. Across the four evaluation sets, Ar+Fr_40k is top on two datasets (TwiFil, WikiFANE). On NER, Ar+Fr_20k yields the best macro-average across WikiFANE, DzNER, and DarNER (64.32), exceeding the strongest baseline (mBERT, 60.32) by 4.00 F1 on average. Aggregating all tasks (TwiFil + NER), Ar+Fr_40k reaches an overall macro-average F1 of 66.07, a +3.24 F1 gain over the best baseline by overall macro-average (AraBERT, 62.83). The largest improvements occur on DzNER (e.g., Ar+Fr_20k: 74.60 vs. MARBERT 67.23). On DarNER, mBERT leads (72.83), while ar-family variants reach 70.28–71.39. In terms of *average rank* across the four datasets (lower is better), table 9 shows that Ar+Fr_40k achieves the best overall rank (1.75); Ar+Fr_20k records 4.75. Among baselines, the best average rank is CAMeLBERT at 7.00.

4.5 Zero-Shot NER Transfer from MSA to Algerian Dialect

We evaluate the zero-shot generalization capability of our smallest model, **Ar_20k**, by adapting it to a named entity recognition (NER) task using only Modern Standard Arabic (MSA) data, and then testing its performance on Algerian dialectal text.

We fine-tune the MSA-adapted Ar_20k model on **ANERCorp**[4], a manually annotated corpus of MSA newswire text, and evaluate it in a zero-shot setting on the DzNER dataset, which consists of Algerian dialectal social media text. This setup allows us to assess the model's ability to generalize across varieties of Arabic without exposure to dialectal data. We adopt this transfer setting because both ANERcorp and DzNER follow the same entity annotation scheme, enabling consistent zero-shot evaluation.

As shown in Table 10, Ar_20k achieves strong performance in the zero-shot setting, outperforming all other models. Among the baselines, only AraBERT v2 surpasses an F1 score of 60, while Ar_20k outperforms both dialect-specific and general-purpose models. Moreover, as shown in Table 7, its performance in the few-shot (supervised fine-tuning) setting rivals models that were explicitly fine-tuned on dialectal data, such as DarijaBERT and DziriBERT. These findings suggest that effective generalization to dialectal NER is possible, even without access to dialectal pretraining data, and may in fact be facilitated by exposure to well-structured MSA during pretraining. This supports the hypothesis that standard Arabic can serve as a robust proxy for dialectal learning when paired with appropriate surface harmonization.

Model	Accuracy	F1
DarijaBERT	91.50	43.96
DziriBERT	92.35	40.53
AraBERT v2	94.33	61.68
CAMeLBERT-DA	87.82	34.73
MARBERT	92.59	53.32
mBERT	90.93	40.96
Ar_20k	**94.55**	**64.16**

Table 10: NER on DzNER in a Zero-Shot MSA Transfer setting

Limitations

Our work presents a novel approach for pretraining dialect-capable Arabic models (Moroccan and Algerian) using only standard language data (MSA and French), guided by phoneme-level normalization. Across both supervised and zero-shot evaluations, our models outperform dialect-pretrained baselines, including in scenarios where no dialectal fine-tuning is used. The consistent vocabulary overlap with dialect-specific models further supports the idea that our subword representations are structurally compatible with North African dialects.

[4] https://huggingface.co/datasets/asas-ai/ANERCorp

While our results are promising, they also underscore a core challenge in working with North African dialects: the severe lack of high-quality, task-diverse benchmarks. Our evaluation is limited to named entity recognition and sentiment analysis, not because of constraints in our approach, but because these are among the few tasks for which annotated data currently exist. Although these tasks offer meaningful insight into the model's generalization abilities, the absence of broader benchmarks for tasks such as question answering, parsing, or text generation restricts our ability to fully evaluate the depth and flexibility of the learned representations. Addressing this gap in resources is essential for advancing dialectal NLP.

Our vocabulary overlap and out-of-vocabulary analyses suggest that phoneme-normalized subwords help support generalization between MSA and dialects. However, we do not include probing-based or contrastive evaluations that could more directly examine representational alignment. Methods such as token-level embedding similarity, attention pattern comparisons, or contrastive alignment tasks may offer additional insight and represent valuable directions for future exploration.

Conclusion

This work introduces a scalable approach for modeling North African Arabic dialects (NADs) without relying on access to dialectal corpora. By pretraining on only Modern Standard Arabic (MSA) and French, augmented with a phoneme-like normalization and vowel-reduction scheme, we achieve strong performance on both sentiment classification and NER tasks across Algerian and Moroccan datasets. These results demonstrate that standard language corpora, when properly normalized and tokenized, can support effective downstream dialect modeling.

Our experiments show that subword-level normalization and data composition choices lead to learned representations that transfer well to dialectal input. Despite being trained exclusively on standard MSA and French language text, the models capture lexical and morphological patterns that generalize across dialectal variation. The low out-of-vocabulary rates and consistent task performance suggest that subword units derived from normalized standard language data are sufficient to support meaningful representation learning, even in the absence of dialect-specific pre-training.

Although we focus on North African Arabic, the general strategy shows promise and could potentially be extended to other dialect-rich languages, though further empirical validation is needed. Future work may explore its application to Gulf Arabic, Maltese, South Asian languages, or other under-resourced spoken varieties.

In sum, we believe our method provides a practical and effective path toward dialect-supportive Arabic models using only standard language corpora, an important step in low-resource NLP for underrepresented varieties.

Acknowledgments

We acknowledge the AID and ANR for their invaluable financial support, which made this research for TRADEF project endeavor possible. The guidance and insights provided were instrumental in the successful execution of our study.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2022. Oscar 22.01: A multilingual dataset of web-scraped text. *arXiv preprint arXiv:2201.06642*.

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, and El Moatez Billah Nagoudi. 2021. 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.

Fahad Alotaibi and Mark Lee. 2013. Automatically developing a fine-grained arabic named entity corpus and gazetteer by utilizing wikipedia. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 392–400.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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.

Abdelhafid 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. *Preprint*, arXiv:1810.04805.

Abdellah El Mekki, Abdelkader El Mahdaouy, Ismail Berrada, and Ahmed Khoumsi. 2021. Domain adaptation for Arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2837, Online. Association for Computational Linguistics.

Khalil Gaanoun, Naira Messaoudi, Ahmed Allak, and Imane Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect. https://huggingface.co/SI2M-Lab/DarijaBERT. SI2M Lab (INSEA Morocco).

Hatem Haddad, Mouna Boussaha, Anis Mahfoudhi, and Lamia Belguith. 2022. Tunbert: The first pre-trained bert model for tunisian arabic. https://huggingface.co/tunis-ai/TunBERT. Instadeep / Tunisia.AI.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.

Salima Harrat, Mourad Abbas, Karima Meftouh, and Kamel Smaïli. 2013. Diacritics Restoration for Arabic Dialects. In *INTERSPEECH 2013 - 14th Annual Conference of the International Speech Communication Association*, Lyon, France. ISCA.

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 (WANLP 2021)*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Khalifa, Muhammad Abdul-Mageed, and Khaled Shaalan. 2021. Self-training pre-trained language models for zero- and few-shot multi-dialectal arabic sequence labeling. *Preprint*, arXiv:2101.04758.

Amina Laggoun, Chahnez Zakaria, and Kamel Smaïli. 2025. Knowledge Distillation for Efficient Algerian Dialect Processing: Training Compact BERT Models with DziriBERT. In *7th International Conference on Advances in Signal Processing and Artificial Intelligence*, Innsbruck (Austria), Austria.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaïli. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, pages 26–34, Shanghai, China.

Lilia Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An algerian corpus and an annotation platform for opinion and emotion analysis (twifil). In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 1202–1210.

Hicham N Moussa and Anas Mourhir. 2023. Darner-corp: An annotated named entity recognition dataset for the moroccan dialect. *Data in Brief*, 48:109234.

Hamdy Mubarak, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. 2019. Highly effective Arabic diacritization using sequence to sequence modeling. 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 2390–2395, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yassine Toughrai, Kamel Smaïli, and David Langlois. 2025. ABDUL: A new approach to build language models for dialects using formal language corpora only. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 16–21, Albuquerque, New Mexico. Association for Computational Linguistics.