ModernAraBERT: Efficient Adaptation of English Language Models for Arabic NLP Tasks

Anonymous ACL submission

Abstract

Transformer-based language models have significantly advanced natural language processing; however, their adaptation to morphologically rich and dialectally diverse languages such as Arabic remains challenging. This paper introduces ModernAraBERT, a resource-efficient adaptation of the English-pretrained ModernBERT model for Arabic NLP tasks. Our approach combines full pretraining on large-scale curated Arabic corpora with lightweight fine-tuning on downstream benchmarks. We evaluate ModernAraBERT across key Arabic NLP tasks, including sentiment analysis, named entity recognition, and extractive question answering, comparing it against strong baselines AraBERTv1,2, MARBERT, and mBERT. Experimental results show that ModernAraBERT consistently outperforms all baselines, achieving substantial gains in Macro-F1 and Exact Match scores. These findings highlight the effectiveness of leveraging modern English architectures for efficient cross-lingual adaptation and provide a scalable pathway for extending state-of-the-art transformer models to Arabic and other underrepresented languages.

1 Introduction

Transformer encoder models such as BERT (Devlin et al., 2019) have transformed natural language processing through their robustness, versatility, and transferability across diverse tasks (Gardazi et al., 2025). Yet, adapting these models to Arabic remains challenging due to the language's rich morphology, complex inflection, and extensive dialectal variation (Matrane et al., 2023).

The rapid progress in Arabic NLP has been driven by the introduction of transformer-based architectures specifically trained or adapted for Arabic. The release of AraBERT (Antoun et al., 2020) represented a major milestone, providing the first large-scale pretrained transformer for Modern Standard Arabic (MSA). Its successor, AraBERTv2 (Antoun et al., 2020), improved on the original through larger and more diverse pretraining data, enhanced vocabulary coverage, and refined preprocessing for Arabic tokenization. Other models extended this direction by addressing dialectal and social-media variations—most notably MARBERT (Abdul-Mageed et al., 2021), which was trained primarily on dialectal Arabic tweets, and multilingual baselines such as mBERT (Devlin et al., 2019), which support Arabic as part of a joint multilingual corpus. While these models have advanced Arabic NLP considerably, they often incur substantial training costs, require specialized data pipelines, and remain limited in efficiently transferring the advances from newer English architectures.

Recent advances in English transformer encoders, particularly ModernBERT (Warner et al., 2024), have introduced architectural and efficiency enhancements that enable large-scale, cost-effective adaptation. Trained on over two trillion tokens, ModernBERT employs rotary positional embeddings (RoPE) (Su et al., 2024), global-local attention, and GeGLU activations (Shazeer, 2020), achieving strong performance with high computational efficiency—making it a robust foundation for adapting to morphologically complex languages such as Arabic.

The main contributions of this paper are as follows.

- We propose a resource-efficient strategy for adapting the English ModernBERT model to Arabic through comprehensive pretraining on curated Arabic corpora, offering a practical alternative to developing new Arabic-specific models from scratch.

- We present ModernAraBERT, an Arabic-adapted variant of ModernBERT optimized for major Arabic NLP tasks.

- We evaluate ModernAraBERT on sentiment analysis, named entity recognition, and question answering, benchmarking it against AraBERTv1, AraBERTv2, MARBERT, and mBERT. The model consistently outperforms all baselines, demonstrating significant accuracy gains for Arabic NLP research and applications.

2 Methodology

2.1 Pretraining Corpora

For the pretraining phase, we compiled a large-scale Arabic corpus from four publicly available sources: OSIAN (Zeroual et al., 2019), the Arabic Billion Words dataset (El-Khair, 2016), the Arabic Wikipedia dump [1], and the OSCAR Arabic dataset (Abadji et al., 2022). These datasets were chosen to ensure coverage of both Modern Standard Arabic and regional variations. The raw texts were preprocessed through a series of normalization steps, including the removal of special characters, punctuation diacritics, elongation characters, and excess whitespace. To further enhance the morphological quality of the data, we used the Farasa segmenter (Abdelali et al., 2016) for affix and root segmentation. The final corpus included over six million sentences, totaling approximately 17 GB in size.

2.2 Tokenization

We extended the original ModernBERT tokenizer, originally trained on English text, by incorporating 80,000 Arabic-specific tokens. The expanded vocabulary includes segmented roots, inflected forms, and common affixes to better capture Arabic morphology and reduce token fragmentation. The 80K size was selected based on coverage analysis, which showed diminishing returns beyond this threshold, balancing representational adequacy and computational efficiency. This choice also aligns with prior Arabic models, where AraBERT employs a 64K vocabulary and MARBERT uses 95K.

2.3 Model Training

Our model is based on the publicly available ModernBERT architecture ModernBERT-base with 22 transformer layers. The embedding layer was resized to accommodate the extended Arabic vocabulary. Pretraining followed the Masked Language Modeling (MLM) objective for three epochs, with sequence lengths of 128 in the first two epochs and 512 in the final epoch to balance efficiency and contextual coverage. The context length was limited to 512 tokens for comparability with baselines like AraBERT. Optimization used AdamW with cosine learning rate decay and gradient clipping. The training progress was tracked via loss and perplexity on a held-out validation set.

The model used in this work is based on the publicly available ModernBERT-base [2] architecture as shown in Figure 1.

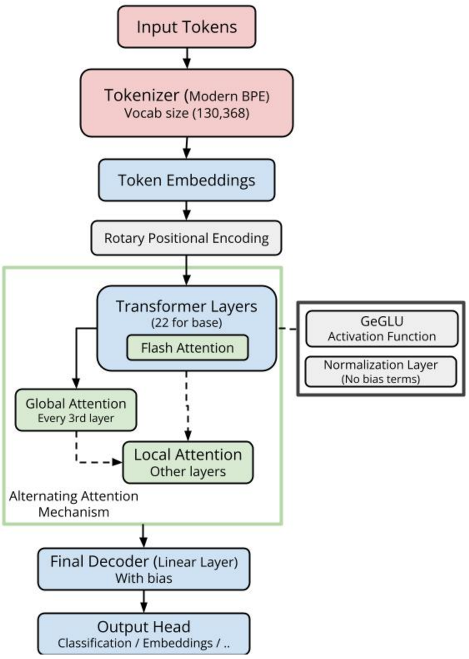

Figure 1: ModernBERT Architecture with extended tokenizer vocabulary size and embedding layer

Our pretrained ModernAraBERT has been made available [3]. Training and benchmark evaluation scripts are also available in our repository [4].

3 Experimental Setup

All pretraining and fine-tuning were conducted on a high-performance node equipped with 12 CPU cores, 32 GB RAM, and a 40 GB NVIDIA A100 GPU. We evaluated ModernAraBERT on three Arabic NLP tasks: Sentiment Analysis (SA), Named Entity Recognition (NER), and Question Answering (QA). During benchmarking, the pretrained encoder was frozen, and only task-specific classification heads were fine-tuned to reduce training time and overfitting on limited datasets. Each task

[1] https://dumps.wikimedia.org/arwiki/

[2] ModernBert-Base

[3] Huggingface URL will be revealed upon review

[4] Our Repository

was trained for up to 200 epochs with an early stopping patience of 10, except NER, which was trained for 5 epochs.

Sentiment Classification We evaluated sentiment classification on three benchmark datasets. The Hotel Arabic Reviews Dataset (HARD) (Elnagar et al., 2018) contains both Modern Standard Arabic (MSA) and dialectal reviews; following (Antoun et al., 2020), neutral 3-star entries were excluded to form a binary classification setup. The Arabic Jordanian General Tweets (AJGT) corpus[5] consists of 1,800 tweets labeled as positive or negative, while the Large-Scale Arabic Book Reviews (LABR) dataset (Aly and Atiya, 2013) was used in its unbalanced binary version for comparability with prior work. For datasets without predefined splits, we adopted a 60/20/20 train/validation/test partition. Sentence-level representations were obtained from the `[CLS]` token and passed through a classification head, with performance evaluated using the Macro F1-score.

Named Entity Recognition NER experiments were conducted on the ANERCorp dataset (Benajiba et al., 2007) using the official CAMeL Lab splits (Obeid et al., 2020), which include entities such as `Person`, `Location`, and `Organization`. The IOB2 tagging scheme (Ramshaw and Marcus, 1999) was employed, assigning the gold entity label to the first subtoken of each word, while continuation subtokens were mapped to the corresponding I-label or masked with `-100` during loss computation. A token classification head was placed on top of the encoder, and results were reported as entity-level macro F1-scores, following standard NER evaluation practice.

Question Answering For extractive question answering, we combined Arabic-SQuAD (Mozannar and Others, 2019) with 50% of ARCD (Mozannar et al., 2019) for training and reserved the remaining ARCD data for testing. The model architecture included the pretrained encoder, a prediction layer, and a linear classifier for start and end span logits. Training employed AdamW (learning rate 3×10^{-5}) with dropout (0.1), a maximum of 200 epochs, and early stopping based on validation F1. Inputs were tokenized to a maximum length of 512 with a document stride of 128, and character-level answer spans were aligned with token indices. Cross-entropy loss was computed jointly over start and end positions, and during inference, the span with the highest joint probability was selected. Evaluation was performed using the Exact Match (EM) metric.

4 Results and Discussion

4.1 Sentiment Classification

Table 1 reports the Macro-F1 scores for sentiment classification across the LABR, HARD, and AJGT datasets. `ModernAraBERT` achieves the best performance on all three benchmarks, surpassing prior Arabic models by a substantial margin. Specifically, it improves over the most widely adopted baseline (`AraBERTv1`) by +11% on LABR, +16% on HARD, and +12% on AJGT. Notably, while `mBERT` and `MARBERT` perform competitively on AJGT, they lag considerably on MSA-heavy datasets such as LABR and HARD. These results indicate that the pretraining of `ModernAraBERT` on large-scale Arabic corpora captures both standard and dialectal features effectively, yielding robust transfer across diverse sentiment domains despite limited fine-tuning.

Table 1: Macro-F1 (%) comparison of `ModernAraBERT` and other Arabic language models across sentiment datasets. Best scores per dataset are in bold.

Model	LABR	HARD	AJGT
AraBERTv1	45.35	72.65	58.01
AraBERTv2	45.79	67.10	53.59
mBERT	44.18	71.70	61.55
MARBERT	45.54	67.39	60.63
ModernAraBERT	**56.45**	**89.37**	**70.54**

4.2 Named Entity Recognition (NER)

Table 2 presents the Macro-F1 results for the NER task. `ModernAraBERT` achieves a clear improvement over all baselines, reaching 28.23% compared to 16.77% for `AraBERTv2` and 13.46% for `AraBERTv1`. This represents a relative gain of over 68% against the strongest prior Arabic model. The performance gap is even larger when compared to `mBERT` and `MARBERT`, which struggle to generalize beyond domain-specific training. These results highlight the ability of `ModernAraBERT` to learn richer token-level representations that effectively capture Arabic morphological and contextual cues, improving entity boundary detection despite limited fine-tuning.

[5] AJGT Dataset

Table 2: Macro-F1 (%) comparison of ModernAraBERT and other models on the NER task. Best score is in bold.

Model	NER (Macro-F1)
AraBERTv1	13.46
AraBERTv2	16.77
mBERT	12.15
MARBERT	7.42
ModernAraBERT	**28.23**

4.3 Question Answering

Table 3 presents the Exact Match (EM) results on the ARCD test split. ModernAraBERT achieves the highest EM score of 27.10%, outperforming all baselines including AraBERTv2 (26.08%), AraBERT (25.36%), mBERT (25.12%), and MARBERT (23.58%). Although the absolute gains appear moderate, they are consistent across all baselines, demonstrating the effectiveness of the adapted model in capturing precise answer spans. The improvements indicate that pretraining on large-scale Arabic corpora enhances contextual understanding and span localization, enabling ModernAraBERT to better align question and context representations in extractive QA tasks.

Table 3: Extractive Question Answering Results (Exact Match, %) on ARCD Test Split.

Model	EM
AraBERT	25.36
AraBERTv2	26.08
mBERT	25.12
MARBERT	23.58
ModernAraBERT	**27.10**

EM: Exact Match.

4.4 Overall Analysis

Across all evaluated NLP tasks, adapting an English-pretrained model like ModernAraBERT to Arabic through our two-phase approach with complete pretraining followed by lightweight fine-tuning of task-specific heads yielded consistently strong performance. The model outperformed AraBERTv1, AraBERTv2, MARBERT, and mBERT baselines, with the most notable gains observed in NER and QA. These results demonstrate the effectiveness of leveraging large-scale English architectures for Arabic transfer learning while maintaining reasonable computational efficiency.

4.5 Hardware Resource Usage

Beyond accuracy, Table 4 summarizes memory consumption across models during head training. AraBERT was the most memory-efficient, showing the lowest RAM and VRAM usage across all tasks, while ModernAraBERT required more resources particularly in QA (5.90 GB RAM, 20.84 GB VRAM) and SA (9.85 GB RAM, 20.63 GB VRAM) due to its larger architecture and extended vocabulary. This increase in computational cost was offset by consistent performance gains, indicating that ModernAraBERT offers a favorable trade-off between accuracy and resource utilization, whereas AraBERT remains better suited for memory-constrained or latency-sensitive scenarios.

Table 4: Hardware Resource Usage Across Models and Benchmarks (Memory usage in GB)

Benchmark	Model	Peak RAM	Peak VRAM
QA	AraBERT	4.52	13.50
	mBERT	1.57	13.66
	MARBERT	1.93	13.60
	ModernAraBERT	5.90	20.84
NER	AraBERT	5.55	6.95
	mBERT	4.85	7.91
	MARBERT	7.76	7.40
	ModernAraBERT	6.49	10.42
SA	AraBERT	8.34	13.50
	mBERT	8.36	13.66
	MARBERT	8.28	13.61
	ModernAraBERT	9.85	20.63

5 Conclusion

In this paper, we introduced ModernAraBERT, a resource-efficient adaptation of the English ModernBERT model for Arabic NLP. By fully pretraining on large-scale Arabic corpora and fine-tuning on downstream tasks, ModernAraBERT achieved consistent gains across sentiment analysis, named entity recognition, and question answering benchmarks. It outperformed strong baselines AraBERT, MARBERT, and mBERT with Macro-F1 improvements exceeding 10% in sentiment analysis, more than doubling NER performance, and yielding the highest Exact Match score in QA. These results demonstrate that modern English architectures can be efficiently transferred to morphologically rich languages, providing a scalable framework for extending state-of-the-art transformer models to Arabic and other underrepresented languages.

6 Limitations

While ModernAraBERT demonstrates consistent improvements across SA, NER, and QA benchmarks, several limitations remain.

First, the adaptation approach focuses on full pretraining followed by head-only fine-tuning. While computationally efficient, this design limits deeper task-specific optimization of the encoder, which may constrain performance in tasks requiring fine-grained reasoning.

Second, the model relies on the Byte-BPE (BBPE) tokenizer used in ModernBERT, which differs from the WordPiece tokenizers employed by most Arabic BERT variants such as AraBERT and MARBERT. Recent findings (Qarah and Alsanoosy, 2024) indicate that BBPE-based models can underperform on extractive QA tasks compared to WordPiece or SentencePiece-based counterparts, particularly when precise span alignment is required. This may explain the relatively smaller performance gain observed for ModernAraBERT on the ARCD dataset compared to its stronger improvements in SA and NER tasks.

Finally, although the experiments cover key Arabic NLP tasks, broader evaluation on additional downstream applications and dialectal datasets would provide a more comprehensive assessment of generalization. Future work will explore tokenizer adaptations and selective layer fine-tuning to further enhance cross-task robustness.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, arXiv:2201.06642.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of NAACL Demonstrations*, pages 11–16.

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.

Mohamed Aly and Amir Atiya. 2013. LABR: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria. 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*.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Ashraf Elnagar, Yasmin S. Khalifa, and Anas Einea. 2018. *Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications*, pages 35–52. Springer International Publishing, Cham.

Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):166.

Yassir Matrane, Faouzia Benabbou, and Nawal Sael. 2023. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 35(6):101570.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Hussein Mozannar and Others. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05685*.

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.

Faisal Qarah and Tawfeeq Alsanoosy. 2024. A comprehensive analysis of various tokenizers for arabic large language models. *Applied Sciences*, 14(13):5696.

L. A. Ramshaw and M. P. Marcus. 1999. Text chunking using transformation-based learning. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 157–176. Springer Netherlands, Dordrecht.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.