Evaluation of Arabic Large Language Models on Moroccan Dialect

Faisal Qarah
Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia
fqarah@taibahu.edu.sa

Tawfeeq Alsanoosy
Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia
tsanoosy@taibahu.edu.sa (corresponding author)

Received: 23 January 2025 | Revised: 28 February 2025 | Accepted: 13 March 2025

ABSTRACT

Large Language Models (LLMs) have shown outstanding performance in many Natural Language Processing (NLP) tasks for high-resource languages, especially English, primarily because most of them were trained on widely available text resources. As a result, many low-resource languages, such as Arabic and African languages and their dialects, are not well studied, raising concerns about whether LLMs can perform fairly across them. Therefore, evaluating the performance of LLMs for low-resource languages and diverse dialects is crucial. This study investigated the performance of LLMs in Moroccan Arabic, a low-resource dialect spoken by approximately 30 million people. The performance of 14 Arabic pre-trained models was evaluated on the Moroccan dialect, employing 11 datasets across various NLP tasks such as text classification, sentiment analysis, and offensive language detection. The evaluation results showed that MARBERTv2 achieved the highest overall average F1-score of 83.47, while the second-best model, DarijaBERT-mix, had an average F1-score of 83.38. These findings provide valuable insights into the effectiveness of current LLMs for low-resource languages, particularly the Moroccan dialect.

Keywords-LLMs; Arabic language; transformers; NLP; Moroccan dialect; distributed computing

I. INTRODUCTION

Large Language Models (LLMs) have shown impressive capabilities in many Natural Language Processing (NLP) tasks in major languages [1, 2]. Most current LLMs have focused mainly on high-resource languages such as English (BERT [3]), French (Camem-BERT [4]), and Spanish (BETO [5]). Even multilingual models, such as mBERT and XLM-R, are restricted to widely used official languages on the Web. In contrast, low-resource languages, including African and Arabic variations, have received comparatively less attention. Arabic serves as a prime example of a low-resource language that encompasses diverse dialects, such as Moroccan Arabic [6], Egyptian Arabic [7], Tunisian Arabic [8], and Gulf Arabic, each with unique linguistic features, grammar, vocabulary, structure, and cultural nuances. These differences indicate that many LLMs might not perform as well in Arabic dialects as they do in English. LLMs often miss important details that make these dialects special, leading to inaccuracies in understanding and generating text. This lack of attention to the unique linguistic and cultural aspects of each dialect can result in low performance. Consequently, some crucial linguistic details may be lost, potentially affecting NLP applications involving Arabic dialects. This gap presents a substantial challenge in the field of NLP. Furthermore, multilingual models often lack the specificity required to accurately represent Arabic dialects such as Moroccan, leading to an additional loss of dialect-specific features. Therefore, increased research and development are imperative to create models that effectively address the linguistic diversity within Arabic.

Arabic is the fourth most widely used language on the Internet, with more than 400 million speakers [9], and is the official language in 22 countries [10]. The Arabic language is highly structured and derivative, with morphology playing a crucial role, and has three primary forms: Modern Standard Arabic (MSA), regional dialects, and Classical Arabic (CA) [9]. However, each country has one or more regional dialects. This variation poses significant challenges for researchers because of the unique structure of the language. Morocco has a variety of languages, primarily dominated by MSA and Moroccan Arabic, known as Darija, being the most prominent. MSA is used in official situations and taught in schools, serving as the formal language for government, media, and literature. In contrast, Darija is a lively everyday language that combines elements of MSA, Amazigh, French, and Spanish. More than 30 million Moroccans use Darija in their daily conversations. While Darija was traditionally an oral language, the rise of social media and increased technological access have facilitated its recent emergence in written form. This newfound written expression is not without challenges, as Darija lacks standardization due to the lack of established grammatical or syntactic rules governing its written use [11]. This variation is further complicated by the fact that Darija can be written using both Arabic and Latin alphabets. Its unique vocabulary and syntax diverge significantly from those of MSA and other languages, presenting substantial challenges to NLP researchers. The Moroccan dialect has few resources, which has led to little attention from the NLP community.

In [12], the performance of ChatGPT in Arabic languages and their dialects was examined. ChatGPT and GPT-4 were compared in Dialectal Arabic (DA) and MSA on tasks related to natural language understanding and generation. A large-scale evaluation included both automated and human assessments, covering 44 distinct language understanding and generation tasks across approximately 70 datasets. The findings showed that smaller models fine-tuned for Arabic consistently outperformed ChatGPT, despite its notable performance in English. Furthermore, this study highlighted the difficulties faced by ChatGPT and GPT-4 in effectively managing Arabic dialects compared to MSA.

In [13], an evaluation study used seven LLMs to measure their performance in Kazakh, a Turkic language. The models were five closed (GPT-3.5, GPT-4, Gemini 1.5 Pro, YandexGPT 2, and YandexGPT 3) and two open (LLAMA 2 and AYA [14]). Six datasets were employed to address different tasks such as answering questions, machine translation, and spelling correction. Overall, the results indicated that the performance of the LLMs on tasks in Kazakh was significantly inferior to that achieved on the corresponding in English. Among the models evaluated, GPT-4 demonstrated the highest performance, followed by Gemini and AYA. The LLMs generally excelled in classification tasks but encountered difficulties with generative tasks. The results highlighted GPT-4 as the most proficient model in the experiment, while Gemini secured the second position in classification tasks. Moreover, AYA showed promising results, particularly given its relatively small size and number of supported languages.

In [15], general models, such as GPT-4 and Llama, were compared with fine-tuned models such as XLM-R, mT5, and Llama-FT in the Urdu language. Both models were compared using a smaller, controlled test set with at least 1,000 samples, covering seven classification tasks. The specialized models consistently outperformed general models in various tasks. The performance of GPT-4 on generation tasks was more in line with human evaluation compared to that of Llama. In [16], GPT-3.5, Llama2-7B, Bloomz-7B1, and Bloomz-3B were compared in a zero-shot setting against SOTA models in Urdu. The SOTA models employed diverse architectures, such as Convolutional Neural Networks (CNN), Random Forests (RF), Sequential Minimal Optimization (SMO), and Long Short-Term Memory (LSTM) with CNN features. The comparison involved 14 tasks and utilized 15 Urdu datasets. The findings showed that SOTA models consistently outperformed LLMs in Urdu NLP tasks under a zero-shot learning paradigm.

Despite extensive evaluations of Arabic language models in various dialects, a significant gap remains in the evaluation of individual Arabic dialects. Existing studies tend to focus on the Arabic language in general, combining MSA and multidialectal datasets in the evaluation process. However, none of these studies have conducted an in-depth evaluation specifically targeting Darija. Furthermore, the unique linguistic features of Darija, which differ significantly from other Arabic dialects, remain underexplored. This gap necessitates a focused investigation into how Arabic LLMs handle the Moroccan dialect across various NLP tasks. The contributions of this paper are as follows:

- This is the first study that evaluates Arabic LLMs in the Moroccan dialect.
- Evaluates 14 language models on 11 datasets tailored for different NLP tasks in the Moroccan dialect, aiming to assist developers and researchers in selecting the appropriate Arabic LLM for use with the Moroccan dialect.

The results of this study provide valuable contributions to the applicability of the currently available LLMs for Arabic dialects. In addition, this study contributes to the methodology of evaluating LLMs and enhancing their quality for mid- and low-resource languages.

II. EXPERIMENTS

All experiments were carried out on a local machine equipped with an AMD Ryzen 9 7950X processor, 64 GB of DDR5 RAM, and two GeForce RTX 4090 GPUs, each with 24 GB of memory. The software environment was established on Ubuntu 22.04, utilizing CUDA 11.8 and the Hugging Face Transformers library [17] to download and fine-tune the comparative language models from the Hugging Face hub.

A. *Fine-tuning Procedures*

To ensure consistency, the same hyperparameters were applied across all experiments to fine-tune the models. Each model was trained with a batch size of 64, a maximum sequence length of 128, and the AdamW optimizer set to a learning rate of 5e-5. For improved training efficiency, mixed precision FP16 was used for gradient computations. All models were evaluated using F1-score and accuracy metrics. For each task, three separate experiments were carried out, each repeated three times. In each experiment, the dataset was split into 80% for training and 20% for validation, using one of the following seeds: 0, 1, and 42. To ensure a fair comparison, the same validation set was utilized across all models within each experiment, and the highest F1-score achieved by each model was reported for each task. The number of epochs and validation steps varied according to the size and complexity of each dataset. Different epoch counts were tested for each task to identify an optimal setting that provided suitable results across all models.

B. *Evaluation Datasets*

The evaluation process involved measuring the performance of the 14 selected LLMs with a total of 11 datasets. Each dataset was used for a single evaluation task, except for OpenDA and DarijaReviews, which were utilized for two different tasks. The total of evaluation tasks was 13. Table I presents a summary of the key characteristics of the datasets included in the study.

TABLE I. KEY CHARACTERISTICS OF THE DATASETS

Datasets	Type	Train size	Validation Size
MYC	SA	15.99k	4k
OMCD	OFD	6.41k	1.60k
MTCD	TC	51.37k	12.84k
MAC	SA	14.46k	3.61k
ElecMor	SA	8.2k	2.05k
MSAC	SA	1.6k	0.4k
MDOD	OFD	16.32k	4.08k
OpenDA-SA	SA	6.81k	1.7k
OpenDA-topic	TC	4.87k	1.21k
MSA_MDA	SA	2.83k	0.7k
TCMD	TC	1.91k	0.48k
DarijaReviews-SA	SA	0.68k	0.17k
DarijaReviews-topic	TC	0.68k	0.17k

SA: Sentiment Analysis, OFD: Offensive Language Detection, TC: Text Classification.

1) Moroccan Youtube Corpus (MYC)

This is a dataset designed for sentiment analysis in the Moroccan dialect, consisting of 20,000 YouTube comments [18, 19] collected from 50 popular Moroccan YouTube channels between 2018 and 2022. The dataset captures comments on various topics, written in both Arabic and Latin. The 20,000 comments were manually labeled with an equal distribution of 10,000 positive and 10,000 negative comments. To address the prevalence of French and English in the collected comments, a cleaning phase removed comments containing only foreign languages while retaining the Latin version.

2) Offensive Moroccan Comments Dataset (OMCD)

This is a dataset specifically designed to detect offensive language in the Moroccan dialect [20, 21]. The dataset comprises comments collected primarily from popular YouTube channels, focusing on trending videos in Morocco. A total of 40,500 comments were initially collected, which were then refined to 8,024 cleaned comments written in the Moroccan dialect. An annotator manually reviewed the comments to exclude those in other dialects or MSA. The annotation process eventually presented 4,304 offensive comments and 3,720 non-offensive comments.

3) Moroccan Topic Classification Dataset (MTCD)

This is the first annotated dataset for topic classification in the Moroccan Arabic language [22, 23]. The dataset comprises 64,222 comments collected from four Moroccan YouTube channels focused on Gaming (13,746), Cooking (10,033), Sports (20,000), and General topics. To compensate for the lack of comments from the Sports channel, additional comments were collected from the Facebook pages of two popular Moroccan soccer teams, Raja and Wydad.

4) Moroccan Arabic Corpus (MAC):

This dataset was designed for sentiment analysis in Moroccan Arabic [24, 25]. Data were extracted from Twitter, resulting in a final dataset of 18,087 valid tweets from an initial collection of 36,114 tweets. The tweets were tagged by two native speakers, with an initial double-labeling process applied to 2,500 tweets. Each tweet was classified according to its polarity and the language used. The labeling process was streamlined using a web application specifically developed for this purpose.

5) Elections Moroccan (ElecMor)

This dataset was designed for sentiment analysis, consisting of 10,254 Arabic Facebook comments [26, 27]. The comments are written in both SA and the Moroccan dialect. The dataset was manually annotated and collected from Facebook, focusing only on political discussions about the 2016 Morocco election. It includes 3,673 positive comments and 6,581 negative comments, making it a mono-topic dataset that exclusively covers political content on social media.

6) Moroccan Sentiment Analysis Corpus (MSAC)

This dataset was designed for sentiment analysis, including reviews and comments from platforms such as Facebook, Twitter, YouTube, and Hespress [28, 29]. It represents a multi-domain corpus that covers a wide vocabulary across sports, social issues, and politics. The final MSAC dataset includes 1,000 positive reviews and 1,000 negative reviews, all written in the Moroccan dialect.

7) Moroccan Darija Offensive Language Detection (MDOD)

This is a human-labeled collection of Moroccan Darija sentences created for offensive language detection [30]. The dataset comprises 20,402 sentences collected from comments on Twitter and YouTube, with 12,685 sentences categorized as non-offensive and 7,717 sentences as offensive. The sentences are presented in both Latin and Arabic scripts.

8) Open Access NLP dataset for Arabic dialects (OpenDA)

This is a collection of datasets gathered from Twitter in various Arabic dialects [31, 32]. The dataset comprises 49,306 posts from five national dialects and was labeled to be used in several NLP tasks, including dialect detection, topic detection, and sentiment analysis. The data was gathered by randomly scraping tweets from active users in a predefined set of Arab countries such as Tunisia, Lebanon, and Morocco. The dataset comprises a total of 9965 Morocco tweets. This dataset was employed for sentiment analysis and text classification tasks for entries labeled Moroccan. The sentiment analysis dataset includes 8,520 posts categorized into three sentiments (positive, negative, and neutral), while the text classification dataset contains 6,091 posts labeled into one of six classes: social, health, politics, sport, economy, and other.

9) Modern Standard Arabic-Moroccan Dialectal Arabic (MSA-MDA)

This sentiment analysis corpus consists of 9,901 comments in MSA and MDA [33, 34], covering various topics relevant to Morocco, including politics, economy, sports, religion, and society, mainly sourced from public Facebook pages of Moroccan news outlets. The dataset comprises 2,221 positive and 4,138 negative comments in MSA, and 684 positive and 2,858 negative comments in MDA. Each comment is annotated to classify its sentiment (positive or negative) and whether it is in MSA or MDA. This study used only MDA comments.

10) Tweet Classification Moroccan Dataset (TCMD)

This dataset was designed for text classification tasks in the Moroccan language [35]. The dataset comprises a total of 2,399 tweets categorized into seven topics: Economy, Culture, Media, International, Politics, Social, and Sports.

11) DarijaReviews

This is a dataset designed for text classification and sentiment analysis tasks in Darija [36]. It consists of reviews of products and services from social media. The dataset comprises a total of 851 positive (456), negative (273), and neutral (122) reviews. Additionally, reviews are labeled into one of 11 topics including cleaning, clothing, cosmetics, entertainment, hospitality, and others.

C. Arabic LLMs

Table II presents a summary of the key characteristics of the pre-trained Arabic models employed.

TABLE II. KEY CHARACTERISTICS OF THE ARABIC PRE-TRAINED MODELS

Model	Number of parameters	Tokenizer	Vocab-size	Dataset size	Dataset type	Dataset sources
AraBERTv0.2-Twitter	136M	WordPiece	64K	77GB	MSA+DA	AraBERTv0.2 + 60M tweets
MARBERTv1	163M	WordPiece	100K	128GB	MSA+DA	1B Arabic tweets
MARBERTv2	163M	WordPiece	100K	198GB	MSA+DA	MARBERTv1 + (unspecified)
CAMeLBERT-DA	108M	WordPiece	30K	54GB	DA	28 different Arabic dialectal corpora
QARiB	136M	WordPiece	64K	97GB	MSA+DA	Arabic Gigaword 5, OBUS, El-Khair, 420M tweets
SaudiBERT	144M	SentencePiece	75K	26.3GB	Saudi DA	STMC + SFC
EgyBERT	144M	SentencePiece	75K	10.4GB	Egyptian DA	ETC + EFC
TunBERT	110M	WordPiece	48.2K	67MB	Tunisian DA	500K Tunisian sentences collected from Common-Crawl
DziriBERT	124M	WordPiece	50K	150MB	Algerian DA	1.2M Algerian tweets
DarijaBERT	147M	WordPiece	80K	691MB	Moroccan DA	(Arabic letters only) Twitter, YouTube, and local Moroccan websites
DarijaBERT-arabizi	170M	WordPiece	110K	287MB	Moroccan DA	(Latinized Arabic only) Twitter, YouTube, and local Moroccan websites
DarijaBERT-mix	209M	WordPiece	160K	1.7GB	Moroccan DA	Twitter, Youtube, and local Moroccan websites
MorRoBERTa	84M	Byte-Level BPE	52K	498MB	Moroccan DA	Facebook, YouTube, and Twitter
MorrBERT	126M	WordPiece	52K	498MB	Moroccan DA	Facebook, YouTube, and Twitter

1) AraBERTv0.2-Twitter

This is a multi-dialect Arabic BERT-based pre-trained language model designed for processing and understanding Arabic text from Twitter [37]. It was built on the original AraBERT model and was furthered pre-trained on a dataset containing 60M DA tweets.

2) MARBERT

MARBERTv1 is a multi-dialect Arabic BERT-based pre-trained language model, specifically targeting diverse Arabic dialects [38]. The model was trained on 1B Arabic tweets from a large internal dataset of approximately 6B tweets. MARBERTv2 is an enhanced version of MARBERT that was pre-trained on the same corpus, in addition to several MSA and the AraNews datasets, featuring an increased sequence length of 512 tokens over 40 epochs.

3) CAMeLBERT-DA

This is a variant of CAMeLBERT, a transformer-based Arabic language model aimed at improving NLP tasks by capturing the complexities of both MSA and dialectal variations [39]. CAMeLBERT-DA was pre-trained using a dataset composed of 54GB of DA text collected from 28 distinct Arabic dialect corpora.

4) QARiB

QARiB is an Arabic model designed for text representation, particularly tailored for tasks involving code-switching and dialectal variations in Arabic [40]. The model was trained on a dataset of approximately 420M tweets and 180M sentences.

5) SaudiBERT

This is a mono-dialect Arabic language model pre-trained exclusively on Saudi dialectal text [41]. The dataset comprises 26.3 GB of text, obtained from the Saudi Tweets Mega Corpus (about 141M tweets) and the Saudi Forum Corpus (about 70M sentences).

6) EgyBERT

EgyBERT is another mono-dialect Arabic language model pre-trained exclusively on Egyptian dialectal texts [7]. The dataset comprises 10.4 GB of text, obtained from two corpora: the Egyptian Tweets Corpus (about 34M tweets) and the Egyptian Forum Corpus (about 44M sentences).

7) TunBERT

This is a mono-dialect Arabic language model specifically fine-tuned for Tunisian Arabic [42]. The model was trained on a 67.2 MB web-scraped dataset, extracted from social media, blogs, and websites, consisting of 500K sentences written in the Tunisian dialect.

8) DziriBERT

This is a mono-dialect Arabic language model pre-trained exclusively on the Algerian dialect [43]. The model was trained on about 1M Algerian tweets.

9) DarijaBERT

DarijaBERT is a mono-dialect Arabic language model pre-trained exclusively Moroccan dialect [22]. The model was trained on a total of 3M Moroccan sentences, both in Arabic and Latin characters, representing 691 MB of text and about

100M tokens. Three different variants exist, and this study used all of them in the evaluation process. DarijaBERT-arabizi was trained in Arabizi text, which is an Arabic Moroccan text written in Latin letters. The dataset was collected from various platforms, including Twitter, YouTube, and local Moroccan websites. DarijaBERT-mix was trained on a larger dataset including both Arabic and Latin characters collected for Twitter, YouTube, and local Moroccan websites.

10)MorrBER

This is a mono-dialect Arabic language model pre-trained exclusively on the Moroccan dialect [44]. MorrBER is structured identically to BERT-base. The model focuses on understanding the linguistic nuances and informal expressions unique to Moroccan dialects. Another variant of MorrBER, called MorRoBERTa, has been proposed, which is a scaled-down variant of the RoBERTa-base model designed for the Moroccan dialect. Both models were trained on a massive corpus of 6M Moroccan dialect sentences collected from Facebook, YouTube, and Twitter, amounting to 71B tokens. This study employed both models.

D. Evaluation Metrics

The evaluation involved measuring the models' performance across all tasks using accuracy and F1-score.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$F1-score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.

III. RESULTS

Table III shows the results of the models in all the datasets used. This table is divided into 3 groups: multi-dialect models, non-Moroccan mono-dialect models, and Moroccan mono-dialect models. All results are rounded to two decimal places. The average score is calculated as the unweighted average of all scores. The highest accuracy or F1-score across all three groups for each dataset is highlighted in bold. In the comparative analysis of the multi-dialect models (AraBERTv0.2-Twitter, QARiB, CAMeLBERT-DA, MARBERTv1, and MARBERTv2), the results show that MARBERTv2 consistently outperformed the others across most datasets, achieving the highest overall average F1-score of 83.47 and accuracy of 88.90, followed by QARiB and MARBERTv1 with average F1-scores of 82.52 and 82.09, respectively. These results suggest that MARBERTv2 demonstrates a strong grasp of the Moroccan dialect, making it particularly effective for tasks that demand high levels of accuracy and consistency in both precision and recall.

TABLE III. EVALUATION RESULTS OF VARIOUS MODELS ON THE MOROCCAN DIALECT TASKS

Model		**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**D13**	**Avg.**
AraBERTv0.2-Twitter	Acc.	87.65	85.92	91.37	90.80	88.35	94.00	90.00	89.32	**93.27**	**93.94**	**85.63**	**85.38**	**76.02**	88.59
	F1	87.61	85.77	91.51	**85.23**	87.00	93.99	89.22	74.58	46.10	89.66	**85.79**	**80.62**	69.49	82.04
QARiB	Acc.	88.52	84.61	91.47	90.24	87.18	95.25	89.81	**91.20**	92.29	93.65	84.58	82.46	73.10	88.03
	F1	88.51	84.51	91.60	84.71	85.86	95.24	89.06	76.72	60.51	89.31	84.51	75.88	66.37	82.52
CAMeLBERT-DA	Acc.	96.95	83.24	91.05	89.25	85.28	93.75	89.37	88.85	92.04	92.24	79.58	80.70	69.01	87.02
	F1	96.95	82.99	91.10	83.37	83.68	93.74	88.66	74.59	61.04	87.16	79.58	69.99	56.96	80.75
MARBERTv1	Acc.	98.70	84.24	90.68	89.83	87.47	95.00	88.95	89.91	91.88	92.67	83.33	78.36	73.68	88.05
	F1	98.70	83.92	90.81	84.01	85.86	95.00	88.28	75.47	59.83	87.53	83.36	71.33	63.03	82.09
MARBERTv2	Acc.	98.50	86.17	91.47	90.74	**88.44**	**95.75**	89.78	90.02	91.06	93.79	85.00	81.87	73.10	88.90
	F1	98.50	86.04	91.62	83.66	**87.12**	**95.74**	89.26	76.79	61.96	89.82	84.97	75.26	64.32	**83.47**
SaudiBERT	Acc.	99.32	85.79	91.74	**90.85**	87.18	94.75	90.05	90.14	92.29	93.79	83.54	82.46	74.27	88.94
	F1	99.32	85.46	91.92	84.72	85.87	94.74	89.44	75.93	57.32	89.32	83.45	75.23	61.98	**82.67**
EgyBERT	Acc.	**99.35**	84.92	90.94	89.36	86.49	93.00	89.54	89.32	91.63	92.52	82.29	76.61	66.67	87.13
	F1	**99.35**	84.77	91.02	82.69	84.98	92.98	88.78	73.61	42.48	87.48	82.40	69.40	45.93	78.91
TunBERT	Acc.	78.67	65.23	72.57	60.25	72.84	73.50	80.30	87.91	92.37	83.22	41.46	65.50	51.46	71.18
	F1	78.59	64.55	73.07	43.93	68.55	73.42	78.12	68.25	40.36	67.38	41.58	56.86	30.43	60.39
DziriBERT	Acc.	98.90	84.74	91.47	87.76	85.52	91.75	90.03	88.91	92.86	92.38	81.67	81.29	73.10	87.72
	F1	98.90	84.64	91.63	80.77	83.89	91.71	89.26	74.74	**65.26**	88.05	81.75	75.53	60.62	82.06
DarijaBERT	Acc.	99.10	85.86	92.07	87.76	85.37	91.50	89.61	89.61	**93.27**	93.65	79.58	81.29	77.19	88.14
	F1	99.10	85.73	92.10	81.67	83.73	91.49	88.92	76.23	48.93	89.24	79.38	75.60	62.22	81.10
DarijaBERT-mix	Acc.	99.15	**87.29**	**92.89**	89.19	86.10	92.75	**90.84**	90.02	91.14	93.65	79.79	**85.38**	**82.46**	89.28
	F1	99.15	**87.22**	**92.95**	82.83	84.61	92.72	**90.26**	75.66	53.30	**90.07**	79.68	79.72	**75.82**	**83.38**
DarijaBERT-arabizi	Acc.	99.12	83.61	91.11	85.49	83.57	89.25	89.24	90.26	92.04	91.54	75.42	83.04	73.10	86.68
	F1	99.12	83.37	91.31	79.12	81.95	89.20	88.46	**77.05**	53.62	85.95	75.65	78.03	55.21	79.85
MorRoBERTa	Acc.	**99.35**	83.36	91.08	85.57	83.81	90.75	89.98	88.67	91.71	91.68	75.00	77.19	71.93	86.16
	F1	**99.35**	83.12	91.12	78.53	81.57	90.73	89.32	74.41	55.90	87.62	75.36	70.25	61.47	79.90
MorrBERT	Acc.	84.35	83.30	90.72	85.52	84.25	90.50	90.76	88.97	91.80	91.96	75.21	81.29	73.10	85.52
	F1	84.35	83.16	90.86	78.58	81.78	90.50	90.02	73.43	57.79	86.77	74.88	76.22	61.18	79.19

D1: MYC , D2: OMCD, D3: MTCD, D4: MAC, D5: ElecMor, D6: MSAC, D7: MDOD, D8: OpenDA-SA, D9: OpenDA-topic, D10: MSA-MDA, D11: TCMD, D12: DarijaReviews-SA, D13: DarijaReviews.

In the comparative analysis of the non-Moroccan mono-dialect models (SaudiBERT, EgyBERT, TunBERT, and DziriBERT), SaudiBERT consistently outperformed the others, achieving an average F1-score of 82.67. This strong performance is likely due to its extensive pre-training on a large Arabic corpus, which enhances its understanding of

Arabic linguistic nuances, including those found in Moroccan dialects. DziriBERT followed closely with an average F1-score of 82.06, likely benefiting from similarities between Algerian and Moroccan dialects. EgyBERT also performed well, with an average F1-score of 78.91, although it trailed behind SaudiBERT and DziriBERT. TunBERT had the lowest average F1-score at 60.39, likely impacted by the smaller dataset it was pre-trained on, despite certain linguistic similarities between the Tunisian and Moroccan dialects compared to Saudi and Egyptian dialects. Future research should explore the factors that contribute to the superior performance of SaudiBERT and EgyBERT over TunBERT, examining both data size and dialectal characteristics.

Finally, in the comparative analysis of the Moroccan dialect models (DarijaBERT, DarijaBERT-mix, DarijaBERT-arabizi, MorRoBERTa, and MorrBERT), the DarijaBERT-mix significantly outperformed the other models on 5 out of 13 tasks, achieving the highest overall average F1-score of 83.38. DarijaBERT ranked second, with the highest scores for both accuracy (88.14) and F1-score (81.10), while MorrBERT achieved the lowest overall average scores. Meanwhile, DarijaBERT-arabizi and MorRoBERTa achieved relatively close results, with F1-scores of 79.90 and 79.85, respectively.

Table IV summarizes the average F1-score of each model across different task groups, specifically sentiment analysis, text classification, and offensive language detection. In the context of sentiment analysis, which included six datasets, namely MYC (D1), MAC (D4), ElecMor (D5), MSAC (D6), OpenDA-SA (D8), MSA-MDA (D10), and DarijaReviews-SA (D12), several models demonstrated strong performance. Among these, MARBERTv2 achieved the highest F1-score of 86.70, followed closely by SaudiBERT at 86.45 and DarijaBERT-mix at 86.39, indicating their effectiveness in capturing sentiment nuances in Arabic Moroccan text. These models showed a clear advantage in handling sentiment analysis tasks compared to others.

TABLE IV. F1-SCORE PERFORMANCE COMPARISON OF VARIOUS MODELS ACROSS DIFFERENT TASK GROUPS.

Model	Sentiment analysis	Text classification	Offensive language detection
AraBERTv0.2-Twitter	85.53	73.22	87.50
QARiB	85.18	**75.75**	86.79
CAMeLBERT-DA	84.21	72.17	85.83
MARBERTv1	85.41	74.26	86.10
MARBERTv2	**86.70**	75.72	87.65
SaudiBERT	86.45	73.67	87.45
EgyBERT	84.36	65.46	86.78
TunBERT	65.28	46.36	71.34
DziriBERT	84.80	74.82	86.95
DarijaBERT	85.29	70.66	87.33
DarijaBERT-mix	86.39	75.44	**88.74**
DarijaBERT-arabizi	84.35	68.95	85.92
MorRoBERTa	83.21	70.96	86.22
MorrBERT	81.66	71.18	86.59

In the text classification task, which included four datasets, namely MTCD (D3), OpenDA-topic (D9), TCMD (D11), and DarijaReviews-topic (D13), QARiB achieved the highest F1-score at 75.75. It was closely followed by MARBERTv2 and DarijaBERT-mix, with F1-score of 75.72 and 75.44, respectively. These results emphasize the effectiveness of QARiB, MARBERTv2, and DarijaBERT-mix in handling text classification in the Moroccan dialect, demonstrating their strong ability to classify diverse content within this task group.

Offensive language detection performance was evaluated using two datasets: OMCD (D2) and MDOD (D7). DarijaBERT-mix demonstrated superior performance with an average F1-score of 88.74, outperforming all others, likely due to its training on a diverse dataset that captures Moroccan cultural and linguistic nuances, especially from social media contexts. Following closely behind, MARBERTv2 achieved an F1-score of 87.65, highlighting its effectiveness but underscoring the specific advantage of DarijaBERT-mix on this task.

Table V presents the top 5 models ranked by overall average F1-score. MARBERTv2 achieved the highest average F1-score (83.47) and the second-highest accuracy (88.90). Although it achieved the highest scores on only two tasks, the model demonstrated consistent performance on all tasks. DarijaBERT-mix obtained the highest accuracy (89.28) and the second-highest F1-score (83.38), outperforming other comparative models on most datasets individually. In particular, SaudiBERT achieved the third-highest scores in both accuracy and F1, likely due to the size and quality of the corpora it was pre-trained on. QARiB followed closely, ranking fourth, while MARBERTv1 ranked fifth in overall average scores. Among all models, TunBERT had the lowest average accuracy and F1-score of 71.18 and 60.39, respectively.

TABLE V. TOP 5 MODELS BY OVERALL AVERAGE F1-SCORE

Rank	Model	Accuracy	F1-score
1	MARBERTv2	88.90%	83.47%
2	DarijaBERT-mix	89.28%	83.38%
3	SaudiBERT	88.94%	82.67%
4	QARiB	88.03%	82.52%
5	MARBERTv1	88.05%	82.09%

IV. CONCLUSION

LLMs demonstrated exceptional performance in a wide range of NLP tasks, particularly for high-resource languages. However, many low-resource languages, such as Arabic and various African languages along with their dialects, remain significantly underrepresented. This raises critical concerns about the ability of LLMs to function equitably in these languages and their dialects. Furthermore, LLM evaluation is of profound significance in the advancement of AI models, which makes it imperative to assess their effectiveness, especially in the context of low-resource languages. Understanding how well LLMs perform in diverse dialects would help ensure inclusivity, diversity, and fairness. This is the first study to evaluate the performance of LLMs in the Moroccan dialect. The performance of several Arabic LLMs was specifically evaluated in Moroccan Arabic, a low-resource dialect spoken by approximately 30 million people. A thorough evaluation of 14 Arabic pre-trained models was performed on the Moroccan dialect, employing 11 datasets across various

NLP tasks. The results showed that MARBERTv2 achieved the highest overall average F1-score of 83.47. In comparison, the second-best model, DarijaBERT-mix, attained an average F1-score of 83.38.

These findings show the importance of evaluating LLMs to improve their performance for low-resource languages. They also highlight opportunities for future research to address current gaps in language representation. Future research should explore techniques to improve the performance of models on diverse Arabic dialects. Also, more in-depth research is needed to understand the factors that contribute to the superior performance of SaudiBERT and EgyBERT compared to TunBERT and DziriBERT. Key aspects to examine include the content of the pre-training corpora, the linguistic similarity between different Arabic dialects, and the impact of using various tokenizers in mitigating the absence of certain dialects from the pre-training data. Additionally, investigating the factors that contribute to the variations in performance across different datasets can provide valuable insights for model development and optimization.

REFERENCES

[1] M. D. Alahmadi, M. Alharbi, A. Tayeb, and M. Alshangiti, "Evaluating Large Language Models' Proficiency in Answering Arabic GAT Exam Questions," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17774–17780, Dec. 2024, https://doi.org/10.48084/etasr.8481.

[2] W. Lei, N. A. S. Abdullah, and S. R. S. Aris, "A Systematic Literature Review on Automatic Sexism Detection in Social Media," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18178–18188, Dec. 2024, https://doi.org/10.48084/etasr.8881.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Minneapolis, MN, USA, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[4] L. Martin *et al.*, "CamemBERT: a Tasty French Language Model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7203–7219, https://doi.org/10.18653/v1/2020.acl-main.645.

[5] J. Cañete, G. Chaperon, R. Fuentes, J. H. Ho, H. Kang, and J. Pérez, "Spanish Pre-trained BERT Model and Evaluation Data." arXiv, Aug. 06, 2023, https://doi.org/10.48550/arXiv.2308.02976.

[6] A. Slim, A. Melouah, U. Faghihi, and K. Sahib, "Improving Neural Machine Translation for Low Resource Algerian Dialect by Transductive Transfer Learning Strategy," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 10411–10418, Aug. 2022, https://doi.org/10.1007/s13369-022-06588-w.

[7] F. Qarah, "EgyBERT: A Large Language Model Pretrained on Egyptian Dialect Corpora." arXiv, Aug. 07, 2024, https://doi.org/10.48550/arXiv.2408.03524.

[8] H. Haddad *et al.*, "TunBERT: Pretraining BERT for Tunisian Dialect Understanding," *SN Computer Science*, vol. 4, no. 2, Feb. 2023, Art. no. 194, https://doi.org/10.1007/s42979-022-01541-y.

[9] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 5, pp. 497–507, Jun. 2021, https://doi.org/10.1016/j.jksuci.2019.02.006.

[10] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, Sep. 2009, https://doi.org/10.1145/1644879.1644881.

[11] M. Ennaji, Multilingualism, Cultural Identity, and Education in Morocco. Springer Science & Business Media, 2005.

[12] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP." arXiv, Oct. 21, 2023, https://doi.org/10.48550/arXiv.2305.14976.

[13] A. Maxutov, A. Myrzakhmet, and P. Braslavski, "Do LLMs Speak Kazakh? A Pilot Evaluation of Seven Models," in *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, Bangkok, Thailand and Online, Dec. 2024, pp. 81–91.

[14] A. Üstün *et al.*, "Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model." arXiv, Feb. 12, 2024, https://doi.org/10.48550/arXiv.2402.07827.

[15] S. Arif, A. H. Azeemi, A. A. Raza, and A. Athar, "Generalists vs. Specialists: Evaluating Large Language Models for Urdu." arXiv, Oct. 03, 2024, https://doi.org/10.48550/arXiv.2407.04459.

[16] M. H. Tahir, S. Shams, L. Fiaz, F. Adeeba, and S. Hussain, "Benchmarking the Performance of Pre-trained LLMs across Urdu NLP Tasks." arXiv, Dec. 31, 2024, https://doi.org/10.48550/arXiv.2405.15453.

[17] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Jul. 2020, pp. 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[18] M. Jbel, M. Jabrane, I. Hafidi, and A. Metrane, "Sentiment analysis dataset in Moroccan dialect: bridging the gap between Arabic and Latin scripted dialect," *Language Resources and Evaluation*, Sep. 2024, https://doi.org/10.1007/s10579-024-09764-6.

[19] J. Mouad, "MouadJb/MYC." Mar. 18, 2025, [Online]. Available: https://github.com/MouadJb/MYC.

[20] K. Essefar, H. Ait Baha, A. El Mahdaouy, A. El Mekki, and I. Berrada, "OMCD: Offensive Moroccan Comments Dataset," *Language Resources and Evaluation*, vol. 57, no. 4, pp. 1745–1765, Dec. 2023, https://doi.org/10.1007/s10579-023-09663-2.

[21] K. Essefar, "kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset." Mar. 07, 2025, [Online]. Available: https://github.com/kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset.

[22] K. Gaanoun, A. M. Naira, A. Allak, and I. Benelallam, "DarijaBERT: a step forward in NLP for the written Moroccan dialect," *International Journal of Data Science and Analytics*, Jan. 2024, https://doi.org/10.1007/s41060-023-00498-2.

[23] "DBert/Data at main AIOXLABS/DBert," *GitHub*. https://github.com/AIOXLABS/DBert/tree/main/Data.

[24] M. Garouani and J. Kharroubi, "MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis," in *Innovations in Smart Cities Applications Volume 5*, 2022, pp. 849–858, https://doi.org/10.1007/978-3-030-94191-8_68.

[25] LeMGarouani, "LeMGarouani/MAC." Mar. 18, 2025, [Online]. Available: https://github.com/LeMGarouani/MAC.

[26] A. Elouardighi, M. Maghfour, and H. Hammia, "Collecting and Processing Arabic Facebook Comments for Sentiment Analysis," in *Model and Data Engineering*, 2017, pp. 262–274, https://doi.org/10.1007/978-3-319-66854-3_20.

[27] "sentiprojects/ElecMorocco2016." Mar. 18, 2025, [Online]. Available: https://github.com/sentiprojects/ElecMorocco2016.

[28] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in *2014 5th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, Apr. 2014, pp. 1–6, https://doi.org/10.1109/IACS.2014.6841964.

[29] "AbderrahmanSkiredj1/MSAC_darija_sentiment_analysis," *Hugging Face*. https://huggingface.co/datasets/AbderrahmanSkiredj1/MSAC_darija_sentiment_analysis.

[30] A. Ibrahimi and A. Mourhir, "Moroccan Darija Offensive Language Detection Dataset." Mendeley Data, Oct. 24, 2023, https://doi.org/10.17632/2y4m97b7dc.2.

[31] E. Boujou, H. Chataoui, A. E. Mekki, S. Benjelloun, I. Chairi, and I. Berrada, "An open access NLP dataset for Arabic dialects : Data

collection, labeling, and model construction." arXiv, Feb. 07, 2021, https://doi.org/10.48550/arXiv.2102.11000.

[32] "Open Datasets - UM6P College of Computing." https://cc.um6p.ma/cc_datasets.

[33] M. Maghfour and A. Elouardighi, "Standard and Dialectal Arabic Text Classification for Sentiment Analysis," in *Model and Data Engineering*, 2018, pp. 282–291, https://doi.org/10.1007/978-3-030-00856-7_18.

[34] "sentiprojects/MSA_MDA_comments." Aug. 07, 2022, [Online]. Available: https://github.com/sentiprojects/MSA_MDA_comments.

[35] O. Lamine, "Tweet_Classification_Moroccan_Dataset." [Online]. Available: https://www.kaggle.com/datasets/omarlamine/tweet-classification-moroccan-dataset.

[36] "ohidaoui/darija-reviews · Datasets at Hugging Face," Nov. 30, 2024. https://huggingface.co/datasets/ohidaoui/darija-reviews.

[37] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding." arXiv, Mar. 07, 2021, https://doi.org/10.48550/arXiv.2003.00104.

[38] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic." arXiv, Jun. 07, 2021, https://doi.org/10.48550/arXiv.2101.01785.

[39] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models." arXiv, Sep. 04, 2021, https://doi.org/10.48550/arXiv.2103.06678.

[40] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations." arXiv, Feb. 21, 2021, https://doi.org/10.48550/arXiv.2102.10684.

[41] F. Qarah, "SaudiBERT: A Large Language Model Pretrained on Saudi Dialect Corpora." arXiv, May 10, 2024, https://doi.org/10.48550/arXiv.2405.06239.

[42] A. Messaoudi *et al.*, "TunBERT: Pretrained Contextualized Text Representation for Tunisian Dialect," in *Intelligent Systems and Pattern Recognition*, 2022, pp. 278–290, https://doi.org/10.1007/978-3-031-08277-1_23.

[43] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, "DziriBERT: a Pre-trained Language Model for the Algerian Dialect." arXiv, Dec. 12, 2022, https://doi.org/10.48550/arXiv.2109.12346.

[44] O. Moussaoui and Y. E. Younnoussi, "Pre-training Two BERT-Like Models for Moroccan Dialect: MorRoBERTa and MorrBERT," *MENDEL*, vol. 29, no. 1, pp. 55–61, Jun. 2023, https://doi.org/10.13164/mendel.2023.1.055.