MagBERT: A Compact Multi-Dialectal BERT for Low-Resource Maghrebi Arabic

Amina Laggoun[1,2,3], Chahnez Zakaria[1], Youness Moukafih[3], Ouassim Karrakchou[3], and Kamel Smaili[2]

[1] École Nationale Supérieure d'Informatique (ESI), Algiers, Algeria
[2] Université de Lorraine, Nancy, France

[3] Université Internationale de Rabat, Rabat, Morocco

Abstract. In a landscape dominated by large language models, Maghrebi Arabic dialects, though widely used in everyday communication and informal writing, remain largely underserved by Natural Language Processing (NLP) technologies. Their limited linguistic resources, high variability, and lack of standardized orthography make them particularly challenging to model effectively. To address these issues, this work introduces MagBERT, a lightweight variant of BERT designed specifically for the three major Maghrebi dialects: Algerian, Moroccan, and Tunisian Arabic, in both Arabic and Latin scripts. The model was pre-trained then fine-tuned on multiple downstream tasks, demonstrating competitive performance compared to several strong benchmark models. Despite its compact size, MagBERT shows strong potential as an efficient and versatile model for processing under-resourced North African dialects.

Keywords: Maghrebi dialects, NLP, Lightweight BERT, Low-resource.

1 Introduction

The three Maghrebi countries; Algeria, Morocco, and Tunisia, share deep historical, cultural, and social ties [1]. Beyond these shared connections, they are also united by the richness and diversity of their local dialects, commonly referred to as Maghrebi Arabic dialects. Although these varieties belong to the broader continuum of Arabic, they differ significantly from one country to another, and even between regions within the same country [2].

These dialects are omnipresent in everyday communication, both in spoken interactions and across social media platforms, forming a central component of the linguistic identity of millions of speakers. In contrast to Modern Standard Arabic (MSA), which enjoys official status and is used in administrative, educational, legal, and media contexts, Maghrebi dialects remain largely non-standardized and non-institutionalized. Their free-form transcription, often involving mixed orthographies such as the use of Latin characters, numerals, or Arabizi transliterations, makes their language modeling particularly challenging. An example of an English sentence translated into MSA and Maghrebi dialects is shown in Table 1. This linguistic complexity is compounded by the scarcity of standardized and annotated resources, which continues to hinder research progress in Natural Language Processing (NLP). As noted by [3], Maghrebi Arabic dialects remain largely underrepresented in NLP applications. Their non-standard nature, high geographic variability, and lack of stable orthographic conventions complicate the development of coherent and reusable textual resources [4]. These constraints not only hinder the creation of large, high-quality corpora but also limit the training of models capable of handling these dialects effectively.

This lack of dedicated resources has become even more critical with the emergence of Large Language Models (LLMs) such as BERT [5], GPT [6], and LLaMA [7]. Although

David C. Wyld et al. (Eds): NLPML, CRIS – 2026
pp. 21-34, 2026. CS & IT -CSCP 2026
DOI: 10.5121/csit.2026.160902

Table 1. Translations of "I haven't slept very well". [1] The Latin-script transcription of MSA is neither common nor officially standardized, yet such informal transliterations can sometimes be encountered on social media.

Language	**Arabic Script**	**Latin Script**
English	-	I haven't slept very well
MSA	لم أنم جيدًا	lam anam jayyidan[1]
Algerian Arabic	ما رقدتش مليح	ma r9dtsh mli7
Moroccan Arabic	ما نعستش مزيان	ma ne3stch mzyan
Tunisian Arabic	ما نمتش بالكداء	ma nemtch blkda

these models have achieved impressive results across a wide range of NLP tasks, their training and deployment require substantial computational, memory, and data resources. This makes them impractical for low-resource or domain-specific contexts. Moreover, the deployment of these models remains challenging on resource-constrained devices such as smartphones, tablets, or IoT systems, which are commonly used in regions with limited connectivity.

That is why, this work focuses on the development of a compact language model MagBERT that handles small amount of data, specifically designed for the three Maghrebi dialects: Algerian, Tunisian, and Moroccan Arabic. MagBERT aims to offer an efficient and balanced representation of Maghrebi dialectal varieties while maintaining a reduced computational footprint.

The remainder of this paper is organized as follows: Section 2 presents the state of the art, reviewing previous research on Arabic and Maghrebi dialects. Section 3 provides an overview of the datasets used in this study, including those employed for the pretraining and fine-tuning of MagBERT. Section 4 details the methodology and contribution, describing the model architecture and pretraining process. Section 5 reports the experimental results and evaluation, followed by a conclusion that summarizes the findings and outlines potential directions for future work.

2 Related Work

Across the Arab world, more than 400 million speakers distributed over 22 countries use the Arabic language in its various linguistic forms [8]. Among these, dialectal Arabic plays a central role in everyday oral communication, often at the expense of Modern Standard Arabic (MSA), which is typically reserved for formal or institutional contexts. This linguistic diversity manifests through a multitude of regional varieties collectively referred to as Arabic dialects. According to [9], from a geolinguistic perspective, Arabic dialects are generally classified into two major zones: the Middle Eastern dialects, which include Egyptian, Sudanese, Levantine (spoken in Syria, Lebanon, Jordan, and Palestine), Peninsular (Gulf), and Mesopotamian (Iraqi) varieties; and the North African dialects (Maghrebi Arabic), which constitute the focus of this study.

Maghrebi dialects are mainly spoken in Algeria, Morocco, Tunisia, Libya, and Mauritania. They differ from Eastern Arabic varieties through distinctive phonetic, morphosyntactic, and lexical features [9]. This work focuses specifically on the dialects of the central Maghreb, namely Algerian, Moroccan, and Tunisian Arabic. These dialects are the outcome of a complex historical evolution shaped by the superposition of multiple linguistic layers over the centuries. Their development reflects a rich and dynamic linguistic coexistence where MSA, dialectal Arabic, Berber languages, French, Spanish and Italian have

long interacted [9].

The daily use of multiple languages in the Maghreb has given rise to a highly frequent linguistic phenomenon: code-switching [10]. This refers to the alternation between two or more languages within the same sentence or utterance [11]. In the Maghrebi context, code-switching often involves a mixture of dialectal Arabic, French, and sometimes other languages. This practice introduces particular challenges in syntactic and semantic analysis [9].

Following the introduction of BERT and its multilingual variant mBERT [12]; including approximately 2% of Arabic data, several models were subsequently trained specifically on MSA, such as AraBERT [13]. The success of AraBERT not only marked a major milestone in Arabic NLP [14], but also revealed its limitations in capturing the rich dialectal diversity inherent to the Arabic language. This realization led to the development of multi-dialectal Arabic models, mostly trained on social media data to better capture the variability and richness of dialectal Arabic. Among these, MarBERT [13], CAMeL-BERT [15], and QARiB [16] stand out, each trained on massive corpora spanning several gigabytes of text. While these models have achieved strong results across various tasks, their broad dialectal coverage has often resulted in reduced specialization, as each Arabic dialect possesses unique phonological, syntactic, and lexical characteristics.

Consequently, researchers have shifted their focus toward mono-dialectal models; each dedicated to a specific dialect. Examples include DziriBERT [17] and its distilled version TinyDziriBERT [18] for Algerian Arabic, DarijaBERT [19] and MorrBERT [20] for Moroccan Arabic, and TunBERT [21] for Tunisian Arabic. Experimental results have consistently shown that these models outperform multi-dialectal ones, even when trained on smaller corpora.

Consequently, researchers have shifted their focus toward mono-dialectal models; each dedicated to a specific dialect. Examples include DziriBERT [17] and its distilled version TinyDziriBERT [18] for Algerian Arabic, DarijaBERT [19] and MorrBERT [20] for Moroccan Arabic, and TunBERT [21] for Tunisian Arabic. Recent works further show that specialized or hybrid architectures continue to outperform generalized models in dialect-specific tasks [22]. Experimental results have consistently shown that these models outperform multi-dialectal ones, even when trained on smaller corpora.

However, all these models rely on the BERT-base architecture, which presents certain limitations. The large number of parameters leads to significant computational costs and long training times. Moreover, deploying such models on edge devices or in real-time systems can be challenging due to their high memory and storage requirements [23, 24]. **With the emergence of large-scale DArabic language models, these efficiency challenges have become even more critical [25]**.

In this context, this work seeks to explore a balance between model size, multi-dialectal coverage, and performance. Specifically, developing a more compact and efficient model that maintains strong performance while effectively representing the linguistic diversity of Maghrebi dialects.

3 Datasets

This section describes the corpus used in the pre-training phase of MagBERT, as well as the diverse corpora used during the fine-tuning phase.

3.1 Pre-training Dataset

MagBERT targets the three Maghrebi dialects; Algerian, Tunisian, and Moroccan, the pre-training corpus comprises data in both Arabic and Latin scripts for each dialect developed in [26].

Table 2 shows the number of tokens issues from each dialect and its proportion in the pre-training corpus.

Table 2. Dialectal composition of the training dataset.

Dialect	**Number of tokens (M)**	**Proportion (%)**
Algerian	23	38
Moroccan	22	36
Tunisian	16	26
Total (Maghrebi)	**61**	**100**

3.2 Fine-tuning Datasets

MagBERT, along with the baseline models, was fine-tuned on multiple datasets for various tasks, including dialect and language identification, topic classification, sentiment analysis and named entity recognition. Each dataset covers either a single dialect or a combination of the three Maghrebi dialects.

Dialect and Language Identification. For this task, four publicly available datasets were used: MADAR [27], Boutef [28], TADI [29].

Multi Arabic Dialect Applications and Resources (MADAR). For this task, only the parallel corpus was used, which contains sentences translated into 25 Arabic dialects, along with English, French, and MSA. For the experiments, only the sentences in Algerian, Tunisian, and Moroccan dialects were retained, all written exclusively in the Arabic script.

In the corpus, Tunisian sentences include two regional variants—Sfax and Tunis, while Moroccan sentences cover Rabat and Fès. Due to the lack of data, to distinguish between local dialects, the variants were grouped into a single entry to have only one dialect per country.

As shown in Table 3, the subset of the MADAR corpus used for training was not balanced across dialects. The models were therefore fine-tuned on both the original unbalanced version (MADAR) and a balanced version (MADAR-bal) to assess the impact of data balance on model performance.

Table 3. Distribution of the number of sentences per dialect in the unbalanced and balanced versions of the filtered MADAR corpus.

Dialect	**MADAR (Sentences)**	**MADAR-bal (Sentences)**
Algerian	2,000	2,000
Tunisian	14,000	2,000
Moroccan	14,000	2,000

Bolstering Our Understanding Through an Elaborated Fake News Corpus (BOUTEF). BOUTEF is a fake news corpus consisting of publications written in multiple languages and dialects, including Modern Standard Arabic (MSA), French, English, the Algerian dialect, and the Tunisian dialect, as well as code-switched content combining two or more of these varieties. The sentences are written in both Arabic and Latin scripts. Table 4 presents the distribution of the selected subset of the BOUTEF corpus in terms of the number of comments per language or dialect.

Table 4. Distribution of comments by language or dialect in the used subset of BOUTEF.

Language / Dialect	**Number of comments**
Algerian Dialect	1078
French	862
MSA	784
Tunisian Dialect	554
Code-switching	421
English	104

Tunisian and Arabic Dialect Identification (TADI). TADI is a binary classification corpus of 50,088 sentences, equally divided between Tunisian dialect (50%) and other Arabic dialects (50%), including Algerian, Egyptian, Jordanian, Libyan, Palestinian, Saudi, Sudanese, Syrian, Emirati, and Yemeni. All sentences are written in the Arabic script. This corpus was used to evaluate the generalization capability of MagBERT when confronted with a broader set of arabic dialects.

Sentiment analysis. For the sentiment analysis task, the models were fine-tuned on the Tunisian Sentiment Analysis Corpus (TSAC) [30] which is a corpus of 17,000 comments in the Tunisian dialect, evenly split between positive and negative sentiment labels. Approximately 1% of the comments are written in the Latin script, while the remainder are in the Arabic script.

Topic classification MagBERT was fine-tuned on the Moroccan Topic Classification Dataset (MTCD) [19]. This corpus consists of 64,000 comments written in the Moroccan dialect and annotated into four topics: Gaming, Kitchen, Sport, and General (mainly covering news and current events). Table 5 presents the distribution of comments across these topics.

Table 5. Distribution of comments in the MTCD corpus.

Topic	**Number of comments**
Gaming	14,000
Kitchen	10,000
Sport	20,000
General	20,000

Named entity recongnition. For this task, the entity-named component of BOUTEF was employed. The entity types along with the distribution of languages in the dataset are summarized in Table 6.

Table 6. Distribution of named entity types and language varieties in the corpus.. PERSON/MISC is a merged category of the entity "person" in the two use cases : the use as a named entity or as qualifying adjective related to a citizenship or belonging to a community.

Category	Label	Description / %
Named Entity Types	LOC	Localisation, place, country, etc.
	ORG	Any organisation
	DATE	Date with different formats
	PERSON/MISC	Person entity or adjectival nationality/community reference
Languages / Dialects	MSA	58.13%
	FRA	20.46%
	ANG	2.54%
	SWITCHING	6.11%
	ALGDIA (Algerian dialect)	8.51%
	ARABIZI	4.26%

4 Contribution

This study introduces MagBERT, a lightweight variant of BERT specifically designed for the three major Maghrebi dialects: Algerian, Moroccan, and Tunisian. The model was initially pre-trained on a multi-dialectal corpus covering these dialects (Section 3) and subsequently fine-tuned on a diverse set of downstream tasks. Its performance was systematically evaluated against several established reference models under identical experimental conditions. To the best of available knowledge, MagBERT is the first BERT-based model in the literature dedicated exclusively to these three dialects.

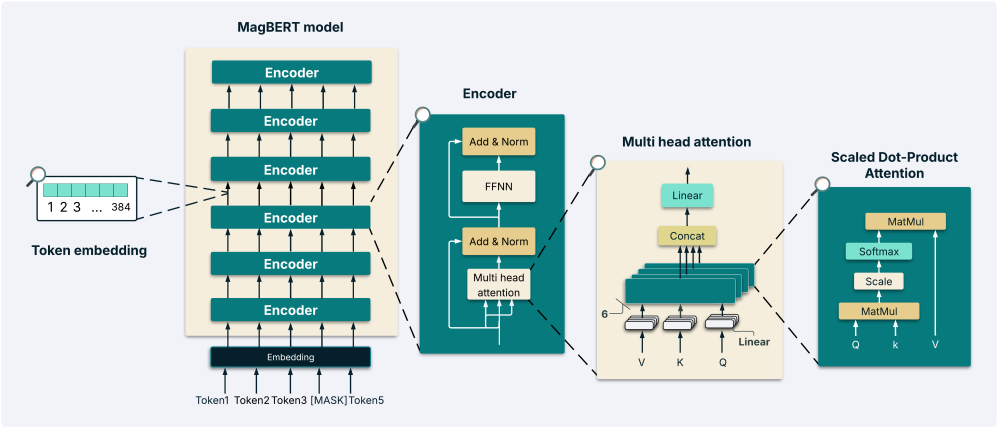

Fig. 1. MagBERT Architecture.

4.1 Model architecture

MagBERT is composed of 6 encoder layers, each with 6 attention heads and a hidden size of 384. This design makes MagBERT a lighter architecture (Figure 1) compared to BERT-base. Table 7 provides a detailed comparison between the two architectures in terms of model depth, number of attention heads, and hidden size.

The choice of the MagBERT architecture is motivated by the growing emphasis on compression techniques within the context of frugal AI. Recent approaches increasingly

Table 7. Comparison between BERT-base and MagBERT architectures.

Model	Layers	Hidden Size	Attention Heads
BERT-base	12	768	12
MagBERT	6	384	6

favor compact models [23], as they enable the design, training, and deployment of systems that require minimal data, energy, and hardware resources, while preserving performance as much as possible.

4.2 Model pre-training

MagBERT was trained using a vocabulary of 70,000 tokens covering the three targeted dialects. This vocabulary was generated from the pre-training corpus using the WordPiece algorithm. Pre-training was performed using the Masked Language Modeling (MLM) objective, where 15% of the tokens in each sentence were masked, and the model was trained to predict them.

Pre-training Configuration

Parameter	Value
Optimizer	Adam
Learning Rate	5×10^{-5}
LR Schedule	Linear decay + 10% warm-up
Epochs	15
Batch Size	16
Gradient Accumulation	Yes
Max Sequence Length	128 tokens
GPU	NVIDIA Tesla V100 (16 GB)
Random Seed	42

Table 8. Pre-training hyperparameters.

Fine-tuning Configuration

Parameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	3–5 (task-dependent)
Early Stopping	Based on validation loss

Table 9. Fine-tuning hyperparameters.

4.3 Model evaluation and baselines

After pre-training, the model was fine-tuned separately for each task on the datasets described in Section 3.2. Fine-tuning was conducted for four epochs, except for the named entity recognition task, which was trained for five epochs. This additional epoch is motivated by the higher complexity of the NER task and the relatively small size of its training dataset. For evaluation, MagBERT was compared with several baseline models

using standard performance metrics. Table 10 compares MagBERT with the baseline models in terms of pre-training corpus size, vocabulary script, number of parameters (#P), and the relative size (*Factor*) of each model with respect to MagBERT.

Table 10. Comparison between MagBERT and baseline models in terms of pre-training corpus size, covered scripts, number of parameters, relative factor with respect to MagBERT (how much larger the models are than MagBERT), and languages/dialects.
*: This is an estimation because this information is missing in the original article of mBERT, A : Arabic, L : Latin, DA : Dialectal Arabic, M : Multi-scripts, ALG : Algerian, MOR : Morrocan, Tun : Tunisian dialects

Model	**Data size**	**Scripts**	**#P(M)**	**Factor**	**Languages/Dialects**
AraBERTv02 [13]	77 GB	A	136	3	MSA
CAMeLBERT-DA [15]	54 GB	A	109	2	MSA + DA
CAMeLBERT-mix [15]	167 GB	A	109	2	MSA + DA
DarijaBERT [19]	691 MB	A	147	3	MOR
DziriBERT [17]	150 MB	A+L	125	3	ALG
MagBERT	**467 MB**	**A+L**	**41**	–	**ALG, MOR, TUN**
MarBERT [14]	128 GB	A	163	4	MSA + DA
mBERT [12]	16* GB	M	167	4	Multiple languages
MorrBERT [20]	700 Mo	A+L	125	3	MOR
QARiB [16]	60 GB	A	135	3	MSA + DA
TinyDziriBERT [18]	178 MB	A+L	18.6	–	ALG
TunBERT [21]	67.2 MB	A	110	2	TUN

Table 10 presents a comparison of MagBERT with five categories of models. The first category includes a multilingual model, represented by mBERT. The second category consists of multidialectal models, such as MarBERT, QARiB, and CAMeLBERT-DA, while the third category corresponds to CAMeLBERT-Mix, trained on a mixture of Modern Standard Arabic (MSA), Classical Arabic, and several dialects. The fourth category comprises monodialectal models, represented by DziriBERT, DarijaBERT, MorrBERT, TunBERT, and TinyDziriBERT. Finally, the fifth category includes AraBERT, trained exclusively on MSA.

MagBERT's vocabulary incorporates both Arabic and Latin scripts, enabling a more faithful representation of North African dialectal usage, particularly on social media. The training data for MagBERT (467 MB) is larger than that of several monodialectal models, such as TunBERT and DziriBERT, reflecting its broader coverage of three Maghrebi dialects. Despite this, all compared models, except TinyDziriBERT, remain two to four times larger than MagBERT in terms of data size or model parameters, highlighting the efficiency of the proposed model.

5 Results

5.1 Dialect and Language Identification

For this task, we conducted experiments on four different corpora (Section 3).

MADAR Results As shown in Table 11, all models achieve relatively high accuracy on the unbalanced version of the MADAR corpus. However, this result should be interpreted with caution, as accuracy is strongly influenced by class imbalance across dialects. In such settings, models may achieve high scores by favoring dominant dialects rather than learning a strong representation of all classes. This is reflected in the gap between accuracy

and F1-score for several models, confirming that F1-score is a more reliable metric under class imbalance.

MagBERT demonstrates strong performance on the unbalanced setting, achieving an F1-score of 88.79%, which is comparable to MarBERT (89.34%) despite the latter being trained on a substantially larger and more diverse corpus. Moreover, MagBERT performs on par with or better than several specialized models such as CAMeLBERT-DA and DarijaBERT, indicating good generalization across dialects. Interestingly, DziriBERT achieves the best performance across both settings. This is somewhat unexpected given that it was trained specifically on the Algerian dialect.

On the balanced version of MADAR, where class distribution is uniform and evaluation is more reliable, MagBERT achieves an accuracy of 87.50% and an F1-score of 87.54%, ranking second overall, just behind DziriBERT. Notably, the performance gap between MagBERT and larger models such as MarBERT narrows or even reverses in this setting, highlighting the efficiency of MagBERT relative to its size.

Table 11. Results on the MADAR corpus for both unbalanced and balanced versions. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	**Factor**	**Unbalanced**		**Balanced**	
		Accuracy (%)	**F1-score (%)**	**Accuracy (%)**	**F1-score (%)**
AraBERTv0.2	3	92.27	86.11	81.25	81.37
CAMeLBERT-DA	2	93.63	88.74	84.67	84.68
CAMel-mix	2	93.32	87.68	87.17	87.22
DarijaBERT	3	93.73	88.83	84.67	84.73
DziriBERT	3	**94.32**	89.91	**87.75**	**87.77**
MagBERT	–	93.73	88.79	87.50	87.54
MarBERT	4	93.90	**89.34**	86.83	86.91
mBERT	4	92.22	86.17	81.5	81.6
MorrBERT	3	93.48	88.05	85.42	85.44
QARiB	3	93.60	88.48	83.92	84.01
TunBERT	2	80.72	67.53	63.17	63.18

Language identification on BOUTEF Table 12 shows the results on the BOUTEF corpus, which includes MSA, English, French, two Maghrebi dialects (Algerian and Tunisian), and code-switching, MagBERT achieved an accuracy of 83.05%. This is only 0.27 percentage points below the best-performing model, DziriBERT (83.32%), despite DziriBERT being nearly three times larger. Furthermore, MagBERT outperforms the distilled version of DziriBERT (79.21%) and several multi-dialectal models, including CAMeLBERT-mix (79.10%), CAMeLBERT-DA (74.77%), and MarBERT (77.92%). These results demonstrate MagBERT's effectiveness in handling multilingual and multidialectal input while maintaining a compact and efficient architecture.

TADI's Results Tests on the TADI corpus involve classifying sentences as Tunisian versus other Arabic dialects covering multiple variants from Algeria, Egypt, Jordan, Libya, Palestine, Saudi Arabia, Sudan, Syria, UAE, and Yemen. MagBERT, despite being trained on only three of these dialects, achieved a competitive accuracy of 93.58%. As shown in Table 13, this performance is comparable to that of large-scale multi-dialectal models trained directly on a wide range of Arabic dialects using gigabytes of dialectal data like CAMeLBERT-mix/da and QARiB, demonstrating MagBERT's ability to capture

Table 12. Performance of models on the Boutef dataset. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	Factor	Accuracy (%)
AraBERTv0.2	3	77.56
CAMeLBERT-DA	2	74.77
CAMeLBERT-mix	2	79.10
DarijaBERT	3	75.43
DziriBERT	3	**83.32**
MagBERT	–	83.05
MarBERT	4	77.92
mBERT	4	71.88
MorrBERT	3	75.56
QARiB	3	77.17
TinyDziriBERT	–	79.21
TunBERT	2	62.94

Tunisian dialectal features. Remarkably, MagBERT also outperforms the Tunisian-specific model TunBERT (81.08%).

Table 13. Performance of models on the TADI dataset. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	Factor	Accuracy (%)
AraBERTv0.2	3	93.24
CAMeLBERT-DA	2	94.35
CAMeLBERT-mix	2	94.54
MagBERT	–	93.58
MarBERT	4	94.60
QARiB	3	**94.68**
TunBERT	2	81.08

5.2 Sentiment analysis

Regarding the sentiment analysis task, MagBERT was evaluated on the TSAC dataset (Table 14), which involves binary classification of Tunisian sentences as positive or negative. MagBERT achieved the highest performance, with an accuracy of 96.68%, outperforming both large-scale multi-dialectal models such as MarBERT (95.61%) and QARiB (96.28%), as well as monodialectal models dedicated to specific varieties, including DarijaBERT, MorrBERT, and even TunBERT (86.67%), which was trained exclusively on Tunisian data.

5.3 Topic classification

In this identification task, the goal is to recognize one of the following Moroccan themes: Gaming, Kitchen, Sport, and General. Table 15 presents the results for this task. MagBERT achieves an accuracy of 90.77% and an F1-score of 90.98% on the MTCD corpus, placing it above several well-known benchmark models, including AraBERTv0.2, CAMeLBERT-mix, CAMeLBERT-DA, and MorrBERT. Although MorrBERT was trained exclusively on Moroccan dialectal data, its performance remains slightly below that of MagBERT, highlighting MagBERT's ability to effectively capture dialect-specific linguistic patterns despite its multi-dialectal pretraining.

Table 14. Performance comparison of models on the TSAC dataset. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	Factor	Accuracy (%)
AraBERTv0.2	3	94.05
CAMeLBERT-DA	2	94.70
CAMeLBERT-mix	2	95.02
DarijaBERT	3	94.76
DziriBERT	3	96.02
MagBERT	–	**96.68**
MarBERT	4	95.61
mBERT	4	86.76
MorrBERT	3	95.02
QARiB	3	96.28
TinyDziriBERT	–	94.52
TunBERT	2	86.67

MagBERT was only outperformed by DarijaBERT and MarBERT. DarijaBERT, nearly three times larger than MagBERT and trained exclusively on Moroccan data, benefits from a stronger specialization in the target dialect. Similarly, MarBERT's superior performance can be attributed to its deeper architecture and larger pretraining corpus. These results underscore the influence of model size and training data scale on overall performance.

Table 15. Performance comparison of models on the MTCD dataset. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	Factor	Accuracy (%)	F1-Score (%)
AraBERTv0.2	3	89.23	89.28
CAMeLBERT-DA	2	89.41	89.56
CAMeLBERT-mix	2	89.02	89.11
DarijaBERT	3	**92.16**	**92.29**
DziriBERT	3	89.56	89.71
MagBERT	–	90.77	90.98
MarBERT	4	91.11	91.15
mBERT	4	88.98	88.76
MorrBERT	3	90.41	90.48
QARiB	3	90.58	90.59
TinyDziriBERT	–	86.65	86.75
TunBERT	2	68.28	68.40

5.4 Named entity recognition

The results presented in Table 16 indicate that, overall, the performance remains relatively low on this multidialectal, multilingual, and multiscript dataset. Nevertheless, it can be observed that AraBERT achieves the highest overall performance on the NER task. This outcome is expected, given that the dataset is predominantly composed of MSA (approximately 60%), which closely aligns with AraBERT's pre-training data.

Similarly, mBERT also performs relatively well, likely due to its multilingual pre-training, which includes portions of MSA, French, and English, languages that are present in the fine-tuning dataset. It is also important to emphasize the relatively strong performance achieved by the small model trained on only 467 MB of text, highlighting its efficiency despite the limited amount of pre-training data and the fact that it has the smallest number

of parameters (41M) among the evaluated models, excluding TinyDziriBERT. Although CAMeLBERT-DA, CAMeLBERT-mix, and QARIB were trained on both MSA and dialectal data and despite the fact that the test dataset consists mainly of MSA, MagBERT achieves comparable or superior results, even when compared to the larger CAMeLBERT models. Furthermore, DziriBERT also achieves lower results than MagBERT.

Overall, these findings suggest that significant efforts are still required to effectively address NER in complex settings involving multidialectal, multilingual, and multiscript named entity corpora.

Table 16. Performance comparison of models on the NER task. Factor represents the relative size compared to MagBERT, i.e., how much larger the models are than MagBERT.

Model	Factor	Precision (%)	F1-Score (%)
AraBERTv0.2	3	**53.33**	**52.52**
CAMeLBERT-DA	2	38.59	41.57
CAMeLBERT-mix	2	46.05	47.86
DarijaBERT	3	39.52	39.52
DziriBERT	3	44.06	47.92
MagBERT	–	46.34	49.11
MarBERT	4	39.81	41.78
MorrBERT	3	42.63	44.15
QARiB	3	45.51	44.89
mBERT	4	50.00	52.05
TunBERT	2	4.25	2.20

6 Conclusion

This work presented MagBERT, a compact transformer-based model specifically designed for the three main Maghrebi dialects: Algerian, Moroccan, and Tunisian Arabic. Through extensive evaluation across several benchmarks, MagBERT demonstrated that it is possible to achieve a practical balance between model size, dialectal coverage, and performance. Despite being significantly smaller than most existing Arabic language models, MagBERT achieves competitive results on multiple tasks, even surpassing larger mono- and multi-dialectal baselines in several cases.

These findings highlight an important insight: efficient models trained on a well-curated mix of dialects can rival larger architectures, proving that linguistic diversity and computational efficiency need not be mutually exclusive. However, the experiments also reveal the limitations of current fine-tuning datasets, which often suffer from imbalance or restricted dialectal variety. Building larger, more balanced, and dialect-aware fine-tuning corpora is essential for more accurate and fair evaluation of future models.

Beyond its empirical results, MagBERT contributes to the advancement of efficient NLP for low-resource and dialect-rich languages. Its lightweight design (41M parameters) makes it particularly well-suited for deployment in resource-constrained environments, such as mobile or embedded systems, where larger models remain impractical. To promote transparency and reproducibility, MagBERT is publicly available on the Hugging Face platform at `https://huggingface.co/NLPinArabIg/MagBERT`

References

1. M. J. Willis, *Politics and Power in the Maghreb: Algeria, Tunisia and Morocco from Independence to the Arab Spring*. London: Hurst Publishers, 2012. [Online]. Available: https://www.hurstpublishers.com/book/politics-and-power-in-the-maghreb/
2. A. Alsudais, W. Alotaibi, and F. Alomary, "Similarities between arabic dialects: Investigating geographical proximity," *CoRR*, vol. abs/2105.04221, 2021. [Online]. Available: https://arxiv.org/abs/2105.04221
3. M. Mashaabi, S. Al-Khalifa, and H. Al-Khalifa, "A survey of large language models for arabic language and its dialects," 2025. [Online]. Available: https://arxiv.org/abs/2410.20238
4. O. Nahli, E. Gugliotta, N. Khlif, and B. Giulia, "Challenges and progress in constructing arabic dialect corpora and linguistic tools: A focus on moroccan and tunisian dialects," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 2023, pp. 293–298.
5. 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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/
6. A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245
7. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971
8. 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, 2021.
9. S. Harrat, K. Meftouh, and K. Smaïli, "Maghrebi Arabic dialect processing: an overview," *Journal of International Science and General Applications*, vol. 1, Mar. 2018. [Online]. Available: https://hal.science/hal-01873779
10. S. Hajbi, Y. Chihab, R. Ed-Dali, and R. Korchiyne, "Natural language processing based approach to overcome arabizi and code switching in social media moroccan dialect," in *Advances in Information, Communication and Cybersecurity*, Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, and A. A. Abd El-Latif, Eds. Cham: Springer International Publishing, 2022, pp. 57–66.
11. K. Darwish, "Arabizi detection and conversion to Arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, N. Habash and S. Vogel, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 217–224. [Online]. Available: https://aclanthology.org/W14-3629/
12. T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" 2019. [Online]. Available: https://arxiv.org/abs/1906.01502
13. W. Antoun, F. Baly, and H. Hajj, "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*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds. Marseille, France: European Language Resource Association, May 2020, pp. 9–15. [Online]. Available: https://aclanthology.org/2020.osact-1.2/
14. M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 7088–7105. [Online]. Available: https://aclanthology.org/2021.acl-long.551/
15. 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," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, and S. Touileb, Eds. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 92–104. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10/
16. A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training bert on arabic tweets: Practical considerations," 2021. [Online]. Available: https://arxiv.org/abs/2102.10684
17. A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, "Dziribert: a pre-trained language model for the algerian dialect," *CoRR*, vol. abs/2109.12346, 2021. [Online]. Available: https://arxiv.org/abs/2109.12346

18. A. Laggoun, C. Zakaria, and K. Smaïli, "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 (ASPAI)*, Innsbruck, Austria, Apr. 2025. [Online]. Available: https://hal.science/hal-04998510f
19. 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*, vol. 9, no. 1, pp. 23–40, Jan. 2024. [Online]. Available: https://doi.org/10.1007/s41060-023-00498-2
20. O. Moussaoui and Y. El Younnoussi, "Pre-training two bert-like models for moroccan dialect: Morroberta and morrbert," *MENDEL*, vol. 29, no. 1, pp. 55–61, Jun. 2023. [Online]. Available: https://mendel-journal.org/index.php/mendel/article/view/223
21. A. Messaoudi, A. Cheikhrouhou, H. Haddad, N. Ferchichi, M. BenHajhmida, A. Korched, M. Naski, F. Ghriss, and A. Kerkeni, "Tunbert: Pretrained contextualized text representation for tunisian dialect," *CoRR*, vol. abs/2111.13138, 2021. [Online]. Available: https://arxiv.org/abs/2111.13138
22. H. Saleh, A. AlMohimeed, R. Hassan, M. M. Ibrahim, S. H. Alsamhi, M. R. Hassan, and S. Mostafa, "Advancing arabic dialect detection with hybrid stacked transformer models," *Frontiers in Human Neuroscience*, vol. 19, p. 1498297, Feb 2025, pMID: 40007884; PMCID: PMC11850318.
23. C. Hu, X. Li, D. Liu, H. Wu, X. Chen, J. Wang, and X. Liu, "Teacher-student architecture for knowledge distillation: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2308.04268
24. J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *CoRR*, vol. abs/2006.05525, 2020. [Online]. Available: https://arxiv.org/abs/2006.05525
25. S. Al-Khalifa, N. Durrani, H. Al-Khalifa, and F. Alam, "The landscape of arabic large language models (allms): A new era for arabic language technology," 2025. [Online]. Available: https://arxiv.org/abs/2506.01340
26. K. Abidi and K. Smaïli, "Creating multi-scripts sentiment analysis lexicons for algerian, moroccan and tunisian dialects," 09 2021.
27. H. Bouamor, N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, and K. Oflazer, "The MADAR Arabic dialect corpus and lexicon," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1535/
28. K. Smaïli, A. Hamza-Jamann, L. David, and A. Djegdjiga, "BOUTEF: Bolstering Our Understanding Through an Elaborated Fake News Corpus," in *The 8th International Conference on Arabic Language Processing*, RABAT, Morocco, Apr. 2024. [Online]. Available: https://hal.science/hal-04578297
29. A. Messaoudi, A. Cheikhrouhou, H. Haddad, N. Ferchichi, M. BenHajhmida, A. Korched, M. Naski, F. Ghriss, and A. Kerkeni, "Tunbert: Pretrained contextualized text representation for tunisian dialect," 2021. [Online]. Available: https://arxiv.org/abs/2111.13138
30. S. Medhaffar, F. Bougares, Y. Estève, and L. Hadrich-Belguith, "Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, N. Habash, M. Diab, K. Darwish, W. El-Hajj, H. Al-Khalifa, H. Bouamor, N. Tomeh, M. El-Haj, and W. Zaghouani, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 55–61. [Online]. Available: https://aclanthology.org/W17-1307/