Language Models and Computational Methods for Low-Resource Languages

Towards Frugal BERT Models for Maghrebi Dialects: Pruning and Distillation Approaches

Amina Laggoun[1,2], **Youness Moukafih**[2], **Ouassim Karrakchou**[2], **and Kamel Smaili**[1]

[1]Université de Lorraine, LORIA, Nancy, France

[2]Université Internationale de Rabat, Rabat, Morocco

{amina.laggoun, kamel.smaili}@loria.fr,
{youness.moukafih, ouassim.karrakchou}@uir.ac.ma

Abstract

The increasing size of pre-trained language models has raised major concerns regarding computational cost, energy consumption, and deployment in resource-constrained settings. These issues are particularly relevant for low-resource and linguistically complex languages such as Maghrebi Arabic dialects, which are highly variable and often involve code-switching. This work explores frugal modeling strategies for Maghrebi Arabic dialects, including pruning, classical knowledge distillation, and dynamic distillation. A set of lightweight and dialect-specific models is developed and evaluated, covering MagBERT, TinyDziriBERT, DynaMoroccanBERT, DynaAlgerianBERT, and DynaTunisianBERT. The analysis focuses on the trade-offs between model size, architectural design, and performance across multiple downstream tasks. Experimental results show that dynamic, task-specific distillation consistently achieves the best balance between efficiency and performance, sometimes slightly outperforming larger baseline models while substantially reducing computational cost, even without additional unlabeled data. Pruning also yields gains, with an improvement of 4.8% in sentiment analysis compared to a larger model. Finally, preserving model depth appears crucial for maintaining strong performance, as Transformer layers are essential for capturing hierarchical representations.

Keywords: Frugal AI, Lightweight BERT, DynaBERT, Arabic dialects.

1 Introduction

Recent advances in natural language processing have been largely driven by large pre-trained language models such as BERT (Devlin et al., 2019), which achieve state-of-the-art performance across a wide range of tasks. However, these models are computationally expensive and memory-intensive, making their deployment challenging in real-world and resource-constrained environments. This has motivated the emergence of frugal AI (Arga et al., 2025), which aims to design efficient models that maintain strong performance while reducing computational cost.

These challenges are particularly pronounced in the context of Arabic dialects, and especially Maghrebi Arabic, which includes the dialects spoken in Algeria, Morocco, and Tunisia. Unlike Modern Standard Arabic, these dialects exhibit substantial linguistic variability, lack standardized orthography, and frequently involve code-switching with languages such as French and English, often written in both Arabic and Latin scripts (Arabizi) (Hamed et al., 2025). In addition, the scarcity of annotated data and the high degree of intra-dialectal variation make the development of robust NLP systems particularly challenging (Abidi and Smaïli, 2021).

To address these issues, several dialect-specific BERT-based models have been proposed, including DziriBERT (Abdaoui et al., 2021), DarijaBERT (Gaanoun et al., 2024), and TunBERT (Messaoudi et al., 2021). While these models achieve strong performance, they remain relatively large, which limits their applicability in resource-constrained settings. More recently, TinyDziriBERT (Laggoun et al., 2025) showed that knowledge distillation can reduce model size while preserving competitive performance.

Despite these advances, the use of frugal modeling strategies for Maghrebi Arabic dialects remains limited. In particular, the comparative impact of different compression approaches and architectural choices on dialectal tasks has not been systematically studied.

This work investigates several complementary frugal strategies for modeling Maghrebi dialects. These include pruning-based compression through

MagBERT, classical knowledge distillation through TinyDziriBERT, and dynamic, task-specific compression using DynaBERT (Hou et al., 2020). The proposed approaches are evaluated on multiple downstream tasks, including dialect and language identification, sentiment analysis, and topic classification, using several datasets.

The main contributions of this work are fourfold. First, we present a comparative study of frugal modeling strategies for Maghrebi Arabic dialects, including pruning, classical distillation, and dynamic distillation. Second, we conduct an empirical evaluation across multiple tasks and datasets, emphasizing the impact of architectural design choices such as width and depth. Third, we show that dynamic, task-specific distillation achieves the best performance–efficiency trade-off, and that preserving model depth is essential for maintaining strong performance. Finally, we provide insights into the influence of teacher quality and cross-dialect generalization in low-resource settings.

The remainder of this paper is organized as follows. Section 2 reviews existing work on model compression and dialectal language modeling. Section 3 presents the proposed approaches, including MagBERT, TinyDziriBERT, and DynaBERT. Section 4 describes the datasets, evaluation tasks, experimental settings, and reports and discusses the experimental results. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Related Work

The growing computational demands of large pretrained language models have motivated extensive research on model compression and frugal AI (Arga et al., 2025), a paradigm that aims to design efficient, resource-aware systems without compromising performance. The main compression approaches include pruning, knowledge distillation, and quantization, all of which seek to reduce the memory and computational footprint of neural models while preserving their ability to generalize across tasks (Kim et al., 2025). These challenges are particularly critical in the context of low-resource languages and dialects, where developing and maintaining large-scale models is both costly and impractical (Arga et al., 2025).

Pruning methods focus on removing redundant parameters or structural components from pretrained models. For instance, LayerDrop (Fan et al., 2019) demonstrated that up to 50% of Transformer layers can be removed with minimal impact on performance. Similarly, CoFi (Xia et al., 2022) proposed a structured pruning approach that jointly removes coarse-grained and fine-grained components, achieving significant speedups while maintaining accuracy.

Knowledge distillation (Hinton et al., 2015) offers an alternative strategy by training a smaller student model to mimic the behavior of a larger teacher model. DistilBERT (Sanh et al., 2020) reduces BERT's size by 40% while retaining most of its performance, whereas TinyBERT (Jiao et al., 2020) further improves compression by distilling intermediate representations and attention mechanisms. More recently, DynaBERT (Hou et al., 2020) introduced a dynamic distillation framework capable of generating multiple sub-networks with varying width and depth, enabling flexible trade-offs between efficiency and performance at inference time.

In parallel, several efforts have focused on developing language models for Arabic dialects, particularly Maghrebi Arabic, which includes the dialects spoken in Algeria, Morocco, and Tunisia (Abdaoui et al., 2021; Messaoudi et al., 2021; Abdul-Mageed et al., 2021; Moussaoui and El Younoussi, 2023; Gaanoun et al., 2024). These dialects present significant challenges for NLP due to their divergence from Modern Standard Arabic (MSA) in vocabulary, morphology, and syntax, as well as their strong influence from Berber, French, and Spanish (Harrat et al., 2018). Additionally, Maghrebi Arabic exhibits frequent code-switching between Arabic and foreign languages, often written in both Arabic and Latin scripts (Arabizi), which further complicates modeling (Hamed et al., 2025). The lack of standardized orthography, combined with limited annotated data and high intra-dialectal variability, makes the development of robust models particularly challenging (Abidi and Smaïli, 2021).

To address these challenges, several dialect-specific BERT models have been proposed. DziriBERT (Abdaoui et al., 2021) was the first model pretrained exclusively on the Algerian dialect, achieving strong performance despite being trained on a relatively small corpus. This was followed by DarijaBERT (Gaanoun et al., 2024) for Moroccan Arabic and TunBERT (Messaoudi et al., 2021) for Tunisian Arabic. However, these models remain relatively large, with parameter counts ranging from 109M to 147M, limiting their applicabil-

ity in resource-constrained environments. More recently, TinyDziriBERT (Laggoun et al., 2025) demonstrated that knowledge distillation can effectively reduce model size while maintaining competitive performance.

Despite these advances, the application of model compression techniques to Maghrebi Arabic dialects remains largely underexplored. This gap is particularly important, as frugal approaches could enable the development of models that achieve performance comparable to existing benchmarks while significantly reducing computational and deployment costs.

3 Models for under-resourced languages

Large Language Models (LLMs) achieve remarkable performance across a wide range of natural language processing tasks. However, training and deploying such models require substantial computational resources, costly hardware infrastructures, and massive amounts of data. Their development also involves extensive training times and significant energy consumption. These constraints raise important economic and environmental concerns, while also limiting the accessibility of state-of-the-art models in resource-constrained settings. This challenge is particularly acute for Maghrebi Arabic dialects, where language resources remain limited and research efforts are still at an early stage. In such contexts, frugal AI offers a promising direction by aiming to design models that are both efficient and competitive while requiring fewer computational resources. To address this issue, several model compression techniques have been explored for Maghrebi Arabic dialects, notably pruning and knowledge distillation. In this work, we investigate three complementary architectures. The first, *MagBERT*, relies on structured pruning by reducing the number of Transformer layers while preserving the original architecture. The second, *TinyDziriBERT* (Laggoun et al., 2025), is a compact student model trained using conventional knowledge distillation, where the student learns from the soft predictions generated by a larger teacher model. The third, *DynaBERT* (Hou et al., 2020), adopts a more flexible distillation strategy by producing dynamic sub-networks with different width and depth configurations. A key distinction lies in the training paradigm. Both MagBERT and TinyDziriBERT require an initial pre-training phase on large unlabeled corpora, followed by task-specific fine-tuning. In contrast, DynaBERT operates directly in a downstream setting: starting from an already pre-trained teacher, it performs distillation during fine-tuning using only the labeled data of the target task. This property makes DynaBERT particularly attractive for low-resource scenarios, where computational efficiency and reduced training costs are essential.

3.1 MagBERT: Maghrebi BERT

MagBERT is a lightweight variant of BERT specifically developed for Maghrebi Arabic dialects, including Moroccan, Algerian, and Tunisian Arabic. Compared with BERT-base, it adopts a significantly smaller architecture with six encoder layers, a hidden size of 384, and six attention heads. This reduced configuration makes MagBERT more suitable for resource-constrained environments while maintaining strong performance on dialectal Arabic tasks. Table 1 summarizes the architectural differences between MagBERT and BERT-base.

Model	Layers	H.Size	Heads
BERT-base	12	768	12
MagBERT	6	384	6

Table 1: Comparison between BERT-base and MagBERT architectures.

MagBERT was first pre-trained on an unlabeled corpus introduced in (Abidi and Smaïli, 2021). This corpus contains sentences in Algerian, Moroccan, and Tunisian dialects, written in both Latin and Arabic scripts. Table 2 reports the token counts for each dialect in the pre-training dataset.

Dialect	Tokens (M)
Algerian	23
Moroccan	22
Tunisian	16
Total	**61**

Table 2: Dialectal composition of the pretraining dataset.

Overall, MagBERT was pre-trained on a relatively modest corpus of 61 million words compared to contemporary large-scale language models. The model employs a WordPiece vocabulary of 70,000 tokens derived directly from the pre-training data. Training follows the Masked Language Modeling (MLM) objective, where 15% of tokens in each sequence are randomly masked. Pre-training was conducted for 15 epochs with a batch size of 16, constrained by the memory limitations of an NVIDIA Tesla V100-PCIE GPU (16 GB). The model was

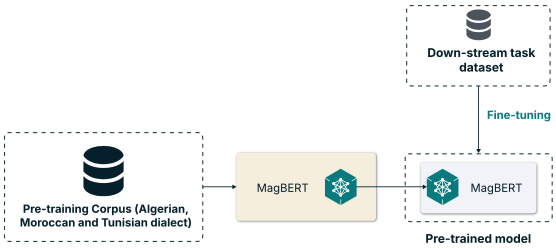

Figure 1: Pre-training and fine-tuning of MagBERT.

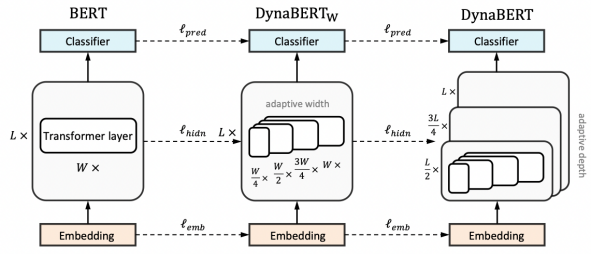

Figure 2: A two-stage procedure to train DynaBERT (Hou et al., 2020).

subsequently fine-tuned on downstream tasks (see figure 1). The design of MagBERT is motivated by the need for an efficient language model tailored to Maghrebi Arabic in low-resource settings. By halving the number of Transformer layers (from 12 to 6), reducing the hidden size (from 768 to 384), and decreasing the number of attention heads (from 12 to 6), the model significantly reduces parameter count and computational cost. This compact design, combined with training on a relatively small dialectal corpus, reflects realistic constraints in under-resourced language settings where large-scale data are unavailable. As a result, MagBERT provides a practical framework for studying the trade-off between model efficiency and downstream task performance in dialectal Arabic processing.

The model was optimized using Adam with a learning rate of 5×10^{-5}, following a linear decay schedule with 10% warm-up, a maximum sequence length of 128 tokens, and gradient accumulation to compensate for the limited batch size. Training used a fixed random seed (42) for reproducibility.

3.2 DynaBERT: DynamicBERT

DynaBERT (Hou et al., 2020) is a distillation framework designed to generate multiple compact sub-models from a single pre-trained BERT model. Its core idea is that one pre-trained and fine-tuned model can produce a family of dynamically adjustable sub-networks. To achieve this, DynaBERT introduces two adaptive dimensions: width and depth. The width corresponds to the number of attention heads and the size of the feed-forward layers, while the depth refers to the number of Transformer layers (i.e., encoder layers in BERT).

As shown in figure 2, the distillation process is carried out in two stages. First, knowledge is transferred from a teacher BERT model to an intermediate model, denoted as DynaBERT$_w$, which supports adaptive width. In this stage, the number of attention heads and feed-forward neurons can be scaled to 25%, 50%, or 75% of the original model size, or kept unchanged. The knowledge transfer leverages the teacher's predictions, embeddings, and hidden representations.

After this stage, multiple width-adaptive configurations are obtained. DynaBERT$_w$ then serves as a teacher assistant for the second stage, where depth adaptability is introduced. Specifically, the number of encoder layers can be reduced to 50% or 75% of the teacher assistant's depth. The final fine-tuning process produces a total of 12 sub-networks with different width–depth trade-offs.

In this work, DziriBERT, DarijaBERT, and TunBERT are used as teacher models. Table 3 summarizes the dialect associated with each teacher model within the DynaBERT framework. The motivation behind this setup is to investigate whether dynamic compression can effectively capture dialect-specific characteristics while leveraging shared representations across sub-networks. Although this approach has shown promising results in English task-specific settings, its effectiveness for dialectal Arabic remains underexplored. This is particularly relevant for Maghrebi dialects, where resources are limited and linguistic variability is high.

Teacher Model	Dialect
DziriBERT	Algerian Arabic
DarijaBERT	Moroccan Arabic
TunBERT	Tunisian Arabic

Table 3: Teacher models and their associated dialects in the DynaBERT framework

3.3 TinyDziriBERT

TinyDziriBERT is trained via knowledge distillation using DziriBERT (Abdaoui et al., 2021) as the teacher model. DziriBERT is a BERT-based pre-trained language model specifically developed for the Algerian dialect, trained on over one million tweets in both Arabic and Latin scripts.

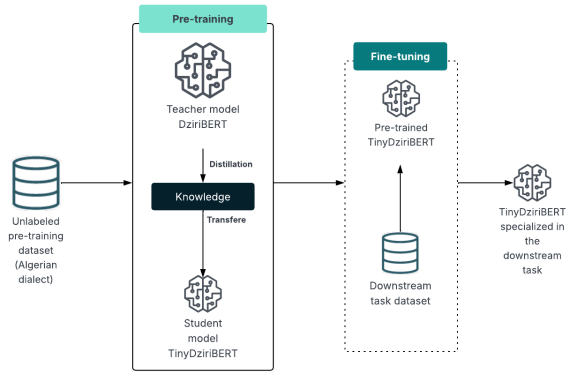

Figure 3: Illustration of the Pre-training and fine-tuning of TinyDziriBERT.

While DziriBERT follows the standard BERT-base architecture, TinyDziriBERT, used as the student model, is significantly smaller in size. Table 4 provides a detailed comparison of their respective architectures. Figure 3 illustrates the process used to obtain a specialized version of the TinyDziriBERT model.

Model	Layers	H.Size	Heads
DziriBERT	12	768	12
TinyDziriBERT	3	256	4

Table 4: Comparison of Teacher and Student Model Architectures.

Both DziriBERT and TinyDziriBERT share the same vocabulary of 50,000 tokens, which corresponds to the teacher's tokenizer.

Since the upper layers of the teacher model capture richer syntactic, semantic, and contextual information, the student model is initialized using a bottom-up layer mapping strategy. Specifically, if the teacher has $N \times k$ layers and the student has Nlayers, the i-th layer of the student is initialized from the $(i \times k)$-th layer of the teacher.

The training process utilizes Masked Language Modeling (MLM), where a 15% of tokens in each input sentence from the dataset are randomly masked. The objective is to predict these masked tokens. The sentences with masked tokens are then processed by both the teacher model and the student model to generate logits.

The student model is then optimized using a hybrid loss function combining Cross-Entropy (CE) loss with respect to the ground-truth labels and Kullback–Leibler (KL) divergence between the teacher and student logits. The total loss is defined as follows:

The total loss is calculated as:

$$\text{Loss} = \alpha \times T^2 \, \text{KL}\left(\frac{\hat{y}_{\text{student}}}{T} \,\middle\|\, \frac{\hat{y}_{\text{teacher}}}{T}\right) + (1-\alpha)\, \text{CE}(y_{\text{student}}, y_{\text{label}}) \quad (1)$$

Where α balances the contribution of the distillation loss and the MLM loss. $\hat{y}_{\text{student}}$ and $\hat{y}_{\text{teacher}}$ are the logits generated by the student and teacher models, respectively, and y_{labels} represents the true tokens. The logits are softened using a temperature parameter T. The KL divergence is scaled by T^2.

During distillation, we used a temperature T = 2 and a distillation coefficient α = 0.5 These values were selected empirically following common practice in BERT distillation.

Table 5 summarizes and compares the models in terms of number of parameters and supported dialects. For DynaBERT, which produces 12 sub-networks across different width and depth configurations, we report the best-performing configuration in terms of the size–performance trade-off. This optimal configuration retains all Transformer layers while reducing the model width to 75%. The sizes of all sub-network configurations are reported in Appendix A.

4 Results and discussions

This section presents the performance of the frugal models: MagBERT, DynaBERT (the DynaBERT sub-network used for comparison in this section has 75% of the original width and the same depth), and TinyDziriBERT, when fine-tuned on several downstream tasks relevant to Maghrebi Arabic dialects. These tasks include dialect and language identification on the MADAR (Multilingual Arabic Dialect Applications and Resources) dataset (Bouamor et al., 2018) and the Boutef (Bolstering Our Understanding Through an Elaborated Fake News Corpus) dataset (Smaïli et al., 2024), sentiment analysis on the TSAC (Tunisian Sentiment Analysis Corpus) dataset (Medhaffar et al., 2017), and topic classification on the MTCD (Maghrebi Topic Classification Dataset) (Gaanoun et al., 2024).

For each task, the models are compared against the top-performing baseline models specific to each dataset and dialect setting.

4.1 Experimental Setup

For all downstream fine-tuning experiments, datasets were split 80%/20% for training and testing. Fine-tuning was conducted for 4 epochs (except NER, 5 epochs, given its task complexity

Model	#P (M)	Scripts	Dialect	U	T
DarijaBERT (Gaanoun et al., 2024)	147	A	Moroccan	✓	✓
DziriBERT (Abdaoui et al., 2021)	124	A+L	Algerian	✓	✓
TunBERT (Messaoudi et al., 2021)	109	A+L	Tunisian	✓	✓
MagBERT	41	A+L	Algerian, Moroccan, Tunisian	✓	✓
TinyDziriBERT (Laggoun et al., 2025)	18	A+L	Algerian	✓	✓
DynaMoroccanBERT	94	A+L	Moroccan	×	✓
DynaAlgerianBERT	77	A+L	Algerian	×	✓
DynaTunisianBERT	66	A+L	Tunisian	×	✓

Table 5: Comparison of Maghrebi Arabic dialect models in terms of number of parameters (#P), vocabulary scripts, supported dialects, and training data requirements. Scripts refer to the writing systems covered in the model's vocabulary: A (Arabic) and L (Latin). U indicates whether the model uses unlabeled data for pre-training, and T indicates whether task-specific data is used for fine-tuning. DynaXBERT denotes a DynaBERT model trained with a dialect X teacher.

and smaller dataset size), using a learning rate of 2×10^{-5} and a batch size of 16, with early stopping based on validation loss.

4.2 Dialect and Language Identification

In this section, we evaluate our models on the language identification task using two distinct datasets.

4.2.1 Language identification on MADAR

The MADAR corpus is a parallel dataset (Bouamor et al., 2018) that contains sentences in 25 Arabic dialects, in addition to Modern Standard Arabic, English, and French. In this study, the focus is only on Algerian, Tunisian, and Moroccan dialects, considering only the Arabic script due to the specific models developed in this study. Due to the limited amount of data available for each sub-dialect, regional variants are grouped together, such as Tunis and Sfax for Tunisian, and Rabat and Fès for Moroccan. The corpus is, however, unbalanced, as shown in Table 6. To assess the effect of data distribution, the models are fine-tuned on both the original unbalanced MADAR dataset and a balanced version (MADAR-bal), in which the dataset size is fixed equally across all classes.

Dialect	MADAR	MADAR-bal
Algerian	2,000	2,000
Tunisian	14,000	2,000
Moroccan	14,000	2,000

Table 6: Distribution of the number of sentences per dialect in the unbalanced and balanced versions of the filtered MADAR corpus.

The results on the MADAR corpus, presented in Table 7, show a consistent trend across all models: performance is generally higher on the unbalanced version of the dataset compared to the balanced one, both in terms of accuracy and F1-score. This can be attributed to the fact that the unbalanced setting preserves the original data distribution, which may implicitly favor dominant dialects present in the training data. This gap suggests that class imbalance in MADAR may inflate accuracy by favoring the dominant Tunisian and Moroccan classes; the balanced setting (MADAR-bal) is therefore a more reliable indicator of true cross-dialect performance, and we recommend it as the primary reference for future comparisons.

Metric	DzB	MagB	TDB	DyAlgB
Factor	–	3	7	2
Acc-Unbalanced	93.16	**93.73**	90.00	93.70
F1-Unbalanced	93.11	88.79	84.80	**93.65**
Acc-Balanced	90.32	87.50	85.10	**90.83**
F1-Balanced	90.39	87.54	84.01	**90.84**

Table 7: Results on the MADAR corpus for both unbalanced and balanced versions. Factor indicates the relative model size with respect to DziriBERT, i.e., how much smaller each model is compared to DziriBERT. DzB: DziriBert, MagB: MagBERT, TDB: TinyDziriBert, DyAlgB: DynaAlgerianBERT

An exception is observed for the smaller models, particularly TinyDziriBERT, which exhibits a noticeable drop in performance, especially on the unbalanced setting, where it significantly underperforms compared to larger architectures such as DziriBERT and DynaBERT. This highlights that highly compact, distillation-based models are particularly sensitive to uneven data distributions.

Among all models, DynaBERT achieves the most competitive and overall best results on both balanced and unbalanced MADAR dataset. Despite reducing the number of attention heads and feed-forward neurons to 75% and having half the size of DziriBERT, it consistently matches or surpasses its performance. This suggests that task-specific dy-

namic distillation is particularly effective for dialect identification.

The full results of all DynaBERT configurations are provided in Appendix B (Tables 15 and 16).

In contrast, MagBERT also performs strongly, outperforming TinyDziriBERT. This can be explained by the fact that TinyDziriBERT relies solely on logit-level distillation from DziriBERT, whereas MagBERT benefits from a more expressive architecture with multiple Transformer layers trained on unlabeled pre-training data.

4.2.2 Language identification on BOUTEF

The tests on this corpus concern BOUTEF which is a datset of fake news collected from social networks. As shown in Table 8, BOUTEF corpus contains sentences in Modern Standard Arabic (MSA), English, French, two Maghrebi dialects (Algerian and Tunisian), as well as code-switching data. Each sentence is labeled according to its corresponding language or dialect.

Table 8: 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

The results on BOUTEF dataset, presented in Table 9, further confirm the effectiveness of DynaBERT in resource-constrained settings.

Model	Factor	Acc	F1
DziriBERT	–	84.16	83.04
MagBERT	3	83.05	81.84
TinyDziriBERT	7	79.21	78.80
DynaAlgerianBERT	2	**84.48**	**83.87**

Table 9: Performance of models on the Boutef dataset. Factor indicates the relative model size with respect to DziriBERT, i.e., how much smaller each model is compared to DziriBERT.

Despite using only 75% of the model width, DynaAlgerianBERT outperforms slightly DziriBERT, which is approximately twice its size, achieving the best performance in terms of both accuracy (84.48%) and F1-score(83.87%). This indicates that task-specific dynamic distillation is able to preserve, and even enhance, discriminative information in multilingual and code-switched contexts, results of the remaining configurations of DynaBERT are provided in Table 17.

In contrast, MagBERT shows performance relatively close to that of DziriBERT despite being three times smaller. This relatively strong performance can be explained by its training strategy, which includes pre-training on the three target dialects (Algerian, Moroccan, and Tunisian). These dialects are known to naturally incorporate lexical borrowing from other languages, particularly French and English, due to historical contact and sociolinguistic influence. Moreover, BOUTEF corpus is extracted from social media data, where users freely mix languages and writing styles without strict linguistic constraints. As a result, MagBERT benefits from exposure to this heterogeneous linguistic environment during pre-training, which improves its robustness in code-switching scenarios.

4.3 Sentiment analysis

Regarding the sentiment analysis task, evaluation is done on TSAC dataset which is a corpus of 17,000 comments in the Tunisian dialect, evenly split between positive and negative sentiment labels. In this case, we used the appropriate LLM TunBERT which is dedicated to Tunisian dialect. DynaBERT was generated from TunBERT to achieve DynaTunisianBERT

Model	Factor	Acc	F1
TunBERT	–	82.67	82.67
MagBERT	3	**96.68**	**96.68**
TinyDziriBERT	6	94.52	94.51
DynaTunisian BERT	2	86.76	86.76

Table 10: Performance comparison of models on the TSAC dataset. Factor represents the relative size compared to TunBERT, i.e., how much smaller the models are than TunBERT.

The results on the TSAC dataset, presented in Table 10, show that MagBERT achieves the best performance, with an accuracy and F1-score of 96.68%, significantly outperforming all other models. This strong performance can be attributed to its pre-training on multiple Maghrebi dialects, which likely improves its ability to generalize to Tunisian sentiment analysis. In contrast, TunBERT, despite being specifically designed for the Tunisian dialect, achieves substantially lower performance. This suggests that the quality and scale of pre-training data play an important role in downstream performance. This also directly impacts the performance of DynaTunisianBERT (more results in 18), which relies

on TunBERT as a teacher model. The relatively low results obtained by DynaBERT highlight a key limitation of knowledge distillation approaches: their strong dependence on the teacher model. When the teacher is suboptimal, the student model inherits these limitations, leading to degraded performance.

Interestingly, TinyDziriBERT achieves competitive results (94.52%), outperforming both TunBERT and DynaBERT. This indicates that cross-dialect transfer is possible within Maghrebi Arabic, even when the student model is trained from a teacher specialized in a different dialect (Algerian). This can be explained by the linguistic proximity between Maghrebi dialects, as well as the availability of a high-quality task-specific dataset such as TSAC, which helps the model adapt effectively during fine-tuning

4.4 Topic classification

In this identification task, the objective is to detect topics expressed in Moroccan dialect and classify them into one of the following categories: Gaming, Kitchen, Sport, and General. The data are sourced from the MTCD corpus. DynaMoroccanBERT is derived from DarijaBERT, a model specifically designed for the Moroccan dialect. The results presented in Table 11, show that DarijaBERT achieves the best performance, with an accuracy and F1-score of 91.99%. DynaMoroccanBERT achieves very competitive results, closely matching DarijaBERT while being approximately twice smaller. This again highlights the effectiveness of task-specific distillation, showing that a well-designed student model can retain most of the teacher's performance even under significant compression. The small performance gap suggests that DynaBERT successfully captures the key task-relevant representations despite reduced size. Results of the other sub-networks of DynaMoroccanBERT are showed in Table 19.

MagBERT also performs well, with results close to the best model, despite being four times smaller than DarijaBERT. Its strong performance can be attributed to its exposure to multiple Maghrebi dialects during pre-training, which likely enhances its ability to generalize across dialectal variations. However, the slight drop compared to DarijaBERT indicates that dialect-specific specialization still provides an advantage for this task.

In contrast, TinyDziriBERT shows a more noticeable decrease in performance. This can be explained by its smaller size and its reliance on a teacher model trained on a different dialect (Algerian), which may limit its ability to fully capture Moroccan-specific thematic patterns.

Model	Factor	Acc	F1
DarijaBERT	–	**91.99**	**91.99**
MagBERT	4	90.77	90.98
TinyDziriBERT	8	87.69	87.68
DynaMoroccanBERT	2	91.82	91.83

Table 11: Performance comparison of models on the MTCD dataset. Factor represents the relative size compared to DarijBERT, i.e., how much smaller the models are than DarijaBERT.

4.5 Error Analysis

To complement the quantitative results, we manually inspected a sample of misclassified instances across tasks. On BOUTEF, most errors involve code-switched sentences where French or English segments dominate the sentence length, causing smaller models (TinyDziriBERT, DynaAlgerianBERT) to default to the majority class. On MADAR-unbalanced, TinyDziriBERT frequently confuses Tunisian and Moroccan sentences sharing common MSA-derived vocabulary, reflecting its narrower Algerian-only pretraining. On TSAC, MagBERT's errors concentrate on short, ambiguous comments (e.g., single-word replies) lacking sufficient context, a limitation shared across all models.

5 Conclusion

This work investigates several frugal AI approaches for Maghrebi Arabic dialects, including pruning and both classical and dynamic knowledge distillation. A set of lightweight and dialect-specific models is introduced, namely MagBERT, TinyDziriBERT, DynaMoroccanBERT, DynaAlgerianBERT, and DynaTunisianBERT. The study analyzes the trade-offs between model size, architectural design, and downstream task performance across multiple benchmarks.

Experimental results demonstrate that dynamic, task-specific distillation consistently achieves the best balance between efficiency and performance. In several cases, it even surpasses larger baseline models while significantly reducing computational cost, without relying on unlabeled data. In particular, DynaBERT configurations with 75% width and full depth consistently provide the most effective size–performance trade-off across most tasks. This confirms that moderate width reduction has

a limited impact on performance, whereas depth reduction is considerably more detrimental.

Beyond architectural aspects, the findings also highlight the importance of pre-training diversity and task-specific adaptation. MagBERT, despite being a layer-pruned model pre-trained on multilingual Maghrebi data, demonstrates strong generalization capabilities, especially on code-switching and cross-dialectal transfer tasks such as BOUTEF and TSAC. This highlights the value of multilingual pre-training even within a closely related language family.

TinyDziriBERT, while the most compact model evaluated, achieves competitive results on several tasks, particularly when cross-dialect transfer within the Maghrebi language family is possible. However, its performance is more sensitive to data imbalance and task complexity, reflecting the limitations of output-level distillation under strong architectural constraints.

Overall, this study shows that frugal NLP approaches are not limited to high-resource languages such as English. Instead, they can be effectively extended to low-resource settings such as Maghrebi dialects, offering promising directions for deploying efficient language models in resource-constrained environments.

Human evaluation was not conducted because the benchmark datasets already provide expert annotations. Establishing human performance on these datasets remains an interesting direction for future work. We report single-run results without variance estimates across seeds; statistical significance testing (e.g., bootstrap resampling) is left for future work.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *CoRR*, abs/2109.12346.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Karima Abidi and Kamel Smaïli. 2021. Creating multiscripts sentiment analysis lexicons for algerian, moroccan and tunisian dialects. In *Proceedings of the 7th International Conference on Data Mining (DTMN 2021)*, Conference Proceedings in Computer Science & Information Technology (CS & IT), Sep 2021, Copenhagen, Denmark, pages 147–156.

Ludovic Arga, François Bélorgey, Arnaud Braud, Romain Carbou, Nathalie Charbonniaud, Catherine Colomes, Lionel Delphin-Poulat, David Excoffier, Christel Fauché, Thomas George, Frédéric Guyard, Thomas Hassan, Quentin Lampin, Vincent Lemaire, Pierre Nodet, Pawel Piotrowski, Krzysztof Sapiejewski, Emilie Sirvent-Hien, and Tamara Tosic. 2025. Frugal ai: Introduction, concepts, development and open questions. *SIGKDD Explor. Newsl.*, 27(1):72–111.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *CoRR*, abs/1909.11556.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. DarijaBERT: a step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*, 9(1):23–40.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched arabic nlp: Progress, challenges, and future directions.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2018. Maghrebi Arabic dialect processing: an overview. *Journal of International Science and General Applications*, 1.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Gun Il Kim, Sunga Hwang, and Beakcheol Jang. 2025. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Comput. Surv.*, 57(10).

Amina Laggoun, Chahnez Zakaria, and Kamel Smaïli. 2025. Knowledge distillation for efficient algerian dialect processing: Training compact BERT models with DziriBERT. In *7th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI)*, Innsbruck, Austria.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

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

Otman Moussaoui and Yacine El Younoussi. 2023. Pretraining two bert-like models for moroccan dialect: Morroberta and morrbert. *MENDEL*, 29:55–61.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Kamel Smaïli, Anissa Hamza-Jamann, Langlois David, and Amazouz Djegdjiga. 2024. BOUTEF: Bolstering Our Understanding Through an Elaborated Fake News Corpus. In *The 8th International Conference on Arabic Language Processing*, RABAT, Morocco.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models.

Appendix A Sub-network sizes of DynaBERT

This appendix reports the size of the DynaBERT sub-network configurations, in terms of the number of parameters, for the three teacher models. The model sizes vary depending on the width and depth multipliers. For example, in the case of DziriBERT, the number of parameters ranges from 12M, corresponding to 25% of the width and 50% of the depth, to 124M for the full model configuration. Here, m_w and m_d denote the width and depth multipliers, respectively.

m_w	m_d	**#P (M)**
0.25	0.50	12
	0.75	14
	1.00	15
0.50	0.50	30
	0.75	36
	1.00	41
0.75	0.50	53
	0.75	65
	1.00	77
1.00	0.50	82
	0.75	103
	1.00	124

Table 12: Sub-network configurations of DynaAlgerianBERT.

m_w	m_d	**#P (M)**
0.25	0.50	18
	0.75	19
	1.00	21
0.50	0.50	42
	0.75	47
	1.00	52
0.75	0.50	71
	0.75	83
	1.00	95
1.00	0.50	105
	0.75	126
	1.00	147

Table 13: Sub-network configurations of DynaMoroccanBERT.

m_w	m_d	#P (M)
0.25	0.50	9
	0.75	10
	1.00	11
0.50	0.50	23
	0.75	28
	1.00	33
0.75	0.50	42
	0.75	54
	1.00	66
1.00	0.50	67
	0.75	88
	1.00	109

Table 14: Sub-network configurations of DynaTunisianBERT.

Appendix B DynaBERT Results

This appendix reports the full evaluation results of all DynaBERT sub-network configurations on the fine-tuning datasets : MADAR, Boutef, TSAC and MTCD.

B.1 DynaBERT results on MADAR

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	87.99	87.98
0.50	0.5	88.41	88.42
0.75	0.5	88.74	88.75
1.00	0.5	89.16	89.17
0.25	0.75	89.07	89.10
0.50	0.75	89.57	89.61
0.75	0.75	90.33	90.37
1.00	0.75	90.16	90.19
0.25	1.0	89.74	89.76
0.50	1.0	90.41	90.43
0.75	1.0	90.83	90.84
1.00	1.0	**91.16**	**91.15**

Table 15: DynaAlgerianBERT ealuation results on the MADAR-Balanced dataset (DziriBERT is the teacher model).

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	93.05	92.99
0.50	0.5	93.20	93.09
0.75	0.5	93.28	93.17
1.00	0.5	93.17	93.07
0.25	0.75	92.98	92.87
0.50	0.75	93.18	93.09
0.75	0.75	93.58	93.44
1.00	0.75	93.65	93.63
0.25	1.0	93.00	92.89
0.50	1.0	93.52	93.42
0.75	1.0	93.70	93.65
1.00	1.0	**93.75**	**93.71**

Table 16: DynaTunisianBERT ealuation results on the MADAR-Unbalanced dataset.

B.2 DynaBERT results on Boutef

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	82.83	81.76
0.50	0.5	83.73	82.80
0.75	0.5	84.17	83.37
1.00	0.5	84.21	83.62
0.25	0.75	83.23	82.34
0.50	0.75	84.08	83.42
0.75	0.75	84.46	83.93
1.00	0.75	84.21	83.59
0.25	1.0	83.35	82.25
0.50	1.0	84.28	83.57
0.75	1.0	84.48	**8**3.87
1.00	1.0	**84.73**	**83.94**

Table 17: DynaAlgerianBERT evaluation results on the Boutef dataset.

B.3 DynaBERT results on TSAC

Width	Depth	Acc (%)	F1 (%)
0.25	0.5	85.82	85.82
0.50	0.5	86.38	86.38
0.75	0.5	86.41	86.41
1.00	0.5	86.50	86.50
0.25	0.75	86.47	86.47
0.50	0.75	86.38	86.37
0.75	0.75	86.38	86.38
1.00	0.75	**86.76**	**86.76**
0.25	1.0	86.56	86.56
0.50	1.0	86.56	86.56
0.75	1.0	86.56	86.56
1.00	1.0	86.68	86.68

Table 18: DynaTunisianBERT evaluation results on the TSAC dataset.

B.4 DynaBERT results on MTCD

Width	Depth	Acc	F1
0.25	0.5	90.85	90.86
0.50	0.5	91.30	91.31
0.75	0.5	91.52	91.52
1.00	0.5	91.63	91.63
0.25	0.75	91.20	91.20
0.50	0.75	91.50	91.50
0.75	0.75	91.64	91.64
1.00	0.75	91.82	91.83
0.25	1.0	91.24	91.24
0.50	1.0	91.67	91.67
0.75	1.0	**91.94**	**91.95**
1.00	1.0	91.92	91.92

Table 19: DynaMoroccanBERT evaluation results on the MTCD dataset.

The detailed results presented in Tables 15–19 provide a comprehensive view of the impact of width and depth scaling on DynaBERT performance across different tasks and datasets. Several consistent trends can be observed.

First, across all datasets, increasing both width and depth generally leads to improved performance, with the best results almost always obtained using the full configuration (width = 1.0, depth = 1.0). However, the performance gains tend to saturate quickly, and smaller configurations remain highly competitive. In particular, configurations with 75% width and full or near-full depth (e.g., 0.75/1.0 or 0.75/0.75) consistently achieve results very close to the full model, while offering a significant reduction in model size. This confirms that a substantial portion of the model redundancy lies in width rather than depth.

Second, depth appears to have a more critical impact on performance than width. Reducing depth (e.g., from 1.0 to 0.5) leads to a more noticeable degradation compared to reducing width. This trend is consistent across MADAR, Boutef, TSAC, and MTCD, suggesting that Transformer layers encode essential hierarchical and contextual representations that are difficult to compress without loss. This observation supports the earlier findings in the main text, where models relying on depth reduction (e.g., MagBERT) showed lower performance compared to DynaBERT configurations that preserve deeper structures.

Third, the effect of compression varies depending on the task. On MADAR and Boutef, which involve dialect and language identification with potentially higher variability, performance differences between configurations are more pronounced. In contrast, on TSAC and MTCD, the performance gap between configurations is relatively small, indicating that these tasks may rely more on surface-level or task-specific features that can be captured even by smaller sub-networks.

The results also reflect the influence of the teacher model. For example, DynaTunisianBERT (TSAC) shows limited variation across configurations, which aligns with the relatively lower performance of its teacher (TunBERT). In contrast, DynaTunisianBERT and DynaMoroccanBERT based on stronger teacher models, exhibit both higher performance and more consistent scaling behavior. This further reinforces the dependency of distillation-based methods on the quality of the teacher model.