

When Does Auxiliary Classification Help Regression in Practice? A Case Study in Arabic Dialectness Modeling

Abdelrahman Sakr, Marwan Torki, and Nagwa El-Makky

Computer and Systems Engineering Department, Alexandria University

{es-abdelrahmansakr2023, mtorki, nagwamakky}@alexu.edu.eg

Abstract

Estimating the Arabic Level of Dialectness (ALDi) is a continuous regression task that quantifies how much a sentence diverges from Modern Standard Arabic (MSA). While auxiliary classification is often used to support regression, its effectiveness depends critically on how the auxiliary task is designed. We conduct a systematic study of auxiliary classification for ALDi estimation, showing that naive auxiliary supervision can degrade regression performance under skewed target distributions. We identify two key factors that determine when auxiliary tasks become beneficial: the granularity of discretization and the choice of loss function. Specifically, 7-bin discretization combined with focal loss substantially improves performance. Our approach achieves an RMSE of 0.1275 on the NADI 2024 blind test set, outperforming previous systems by over 9%, and the same design transfers to English toxicity regression. Detailed error analysis reveals how auxiliary classification reshapes model behavior across the Arabic dialect spectrum. We release our code and model predictions to facilitate future research.¹

1 Introduction

Arabic is a diglossic language: Modern Standard Arabic (MSA) governs formal writing, while a rich continuum of regional dialects shapes everyday speech and social media. These varieties differ substantially in morphology, syntax, and vocabulary (Habash, 2010). Critically, dialectal variation is *not binary*—a sentence can blend MSA and colloquial features at any ratio—so reducing it to a discrete label discards linguistically meaningful gradations.

Arabic Level of Dialectness (ALDi) estimation formalizes this by assigning each sentence a continuous score in $[0, 1]$, where 0 denotes pure MSA

and 1 indicates maximal dialectal divergence (Keleg et al., 2023). Yet continuous regression over this spectrum is fundamentally challenging: the AOC-ALDi (Keleg et al., 2023) training distribution is heavily skewed toward MSA (Figure 1), causing standard regression models to collapse toward conservative mean predictions and systematically under-score high-dialect sentences.

The natural remedy is to enrich the regression objective with an *auxiliary classification signal*—jointly training the shared encoder on the regression task and an auxiliary classification task, so its representations benefit from both signals—a well-known multi-task strategy for regularizing shared encoders (Caruana, 1997; Liu et al., 2019). However, our central finding is that **auxiliary classification is not inherently beneficial**: naively adding a standard Cross-Entropy head *degrades* regression performance—a concrete instance of the “negative transfer” problem (Jiang et al., 2023) where auxiliary gradients conflict with the primary objective. The real question is therefore not *whether* to add an auxiliary task, but *how* to design it.

We answer this through a systematic staged study on the NADI 2024 ALDi benchmark (Abdul-Mageed et al., 2024), identifying two critical design factors: (1) **discretization granularity**—bins must be rich enough for informative supervision yet coarse enough to avoid data sparsity; and (2) **loss function choice**—Focal Loss (Lin et al., 2020) redirects the auxiliary signal away from MSA-heavy easy examples toward boundary-adjacent ambiguous cases where guidance is most needed. A controlled ablation confirms these are architectural contributions, not artefacts of hyperparameter tuning, and the same ablation holds on English toxicity regression. Together, 7-bin discretization and Focal Loss transform a harmful auxiliary head into one that substantially improves regression. Our best configuration uses 7 bins, Focal Loss with $\gamma = 3$, and a regression loss weight

¹<https://github.com/Body123/ALDi-Arabic-Dialectness-MTL>

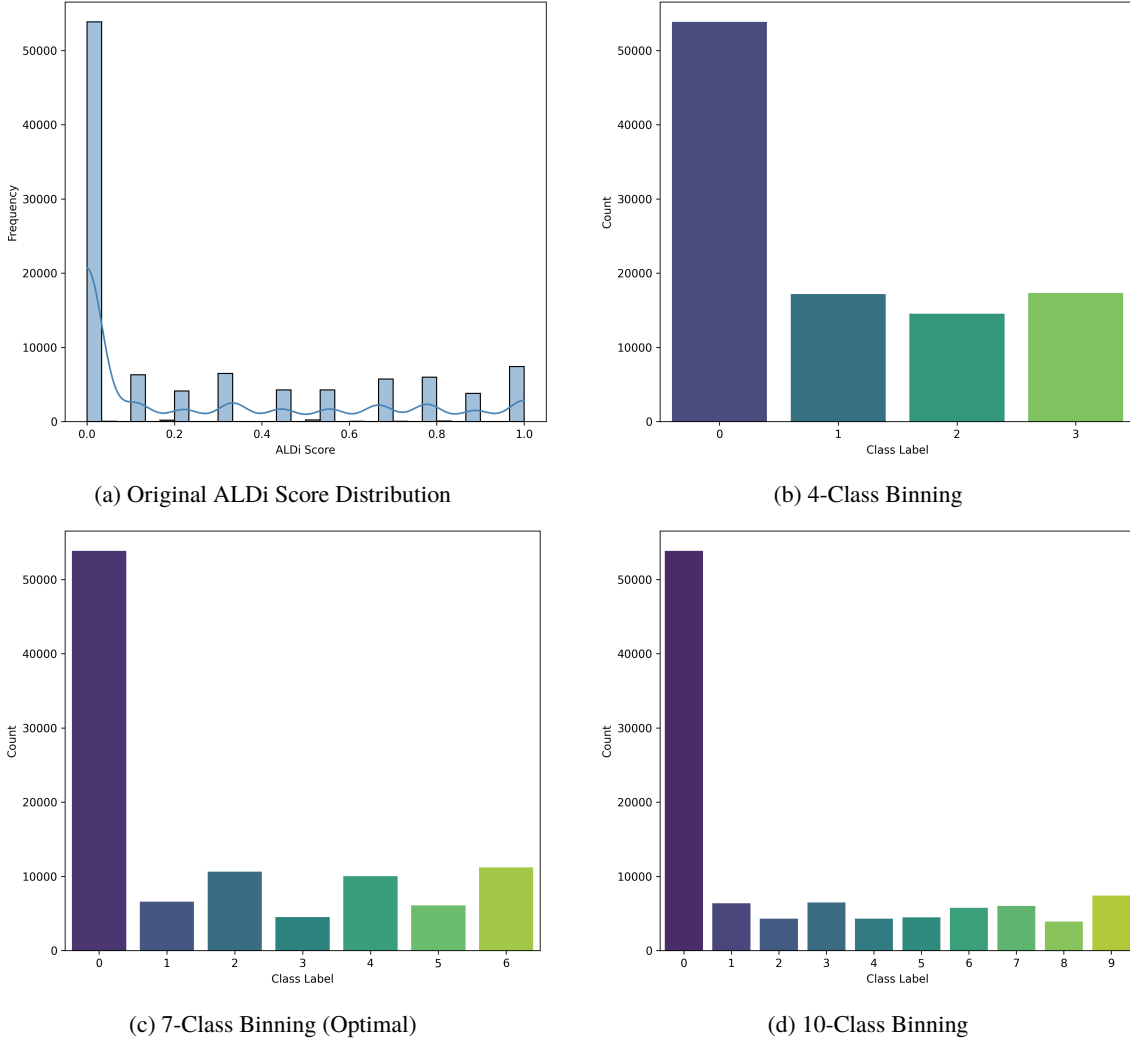


Figure 1: Comparison of binning strategies. (a) The original distribution is skewed towards MSA. (b) 4-class binning is balanced but coarse. (c) 7-class binning offers the best trade-off between signal richness and class density. (d) 10-class binning is too granular, causing data sparsity.

of $\alpha_{\text{reg}} = 0.73$. It achieves a blind-test Root Mean Squared Error (RMSE) of **0.1275**—surpassing the previous best (0.1403) by over 9%.

2 Related Work

Arabic Dialect Modeling: From Discrete Labels to Continuous Spectra. Early Arabic NLP framed dialectness as categorical, tackling dialect identification at the sentence (Zaidan and Callison-Burch, 2011; Salameh et al., 2018) and token levels (Solorio et al., 2014; Molina et al., 2016). This discrete framing persisted through shared tasks such as NADI 2022, which posed dialect identification as a classification problem (Abdul-Mageed et al., 2022). Keleg et al. (2023) challenged this paradigm by introducing the AOC-ALDi dataset and showing that discrete classifiers fail to capture

the *gradual* nature of dialectal variation, motivating a shift to continuous regression. NADI 2024 Subtask 2 (Abdul-Mageed et al., 2024) operationalized this shift as a shared task. Prior submissions addressed the resulting regression challenge using data-level strategies such as data augmentation (Nacar et al., 2024) and regression ensembles with contrastive learning, weighted random sampling, and undersampling of the MSA-dominated majority class (Sakr et al., 2024). However, neither investigated *how* the regression objective itself should be structured under a skewed distribution. In contrast, our approach leaves the training distribution untouched. More broadly, skewed continuous targets are studied as *imbalanced regression*, with remedies spanning data-level resampling (Branco et al., 2019), loss reweighting, and representation-level smoothing (Yang et al., 2021;

Puetz et al., 2026).

Multi-Task Learning and the Problem of Negative Transfer. Multi-task learning improves generalization by using related tasks as inductive biases (Caruana, 1997; Liu et al., 2019; Chen et al., 2024), and has benefited a range of applications in NLP (Fadel et al., 2024) and computer vision (Foglia et al., 2023). However, a well-documented failure mode—*negative transfer*—occurs when auxiliary gradients conflict with the primary objective, degrading shared representations for both tasks (Jiang et al., 2023). This failure is especially likely when the auxiliary signal is dominated by easy examples, providing no useful information and actively misleading the encoder. This is precisely what we observe in Stage 2 of our experiments. While prior remedies focus on gradient surgery and dynamic loss reweighting (Jiang et al., 2023), we show that *loss function design for the auxiliary head itself* is an equally important and more accessible lever.

Regression via Classification and Discretization Design. Discretizing continuous targets into classification bins is widely used in depth estimation and age estimation, where it often outperforms direct regression by providing a more stable learning signal (Berg et al., 2021). Pintea et al. (2023) further find this benefit to be strongest when the target distribution is imbalanced. The same strategy has been carried into imbalanced regression specifically, where the standard benchmark suite includes text: STS-B-DIR (Yang et al., 2021) pairs sentences with continuous, skewed similarity scores—the same regime of skewed continuous supervision we face—and Pu et al. (2025) show from a Bayesian view that the imbalanced-regression objective decomposes into classification plus regression, improving both facial-age and STS-B-DIR prediction. However, these gains hinge on *how* the target space is discretized: converting a continuous target into discrete classes discards ordinal information and introduces discretization error (Zhang et al., 2023), while overly fine bins cause class sparsity. Critically, the auxiliary loss function then determines *which* examples drive learning—Cross-Entropy weights all bins equally and is biased toward the majority class, while Focal Loss (Lin et al., 2020) concentrates the signal on hard, boundary-adjacent instances. Each factor has been studied in isolation; we are the first to study the *interaction* of discretization granular-

ity and auxiliary loss design for a continuous linguistic spectrum task, grounding both choices in a principled staged analysis.

3 Proposed Framework

3.1 Architecture

We formulate Arabic Level of Dialectness (ALDi) estimation as a multi-task learning (MTL) problem (Liu et al., 2019), shown in Figure 2. We adopt AraBERTv2-Twitter² as the shared encoder, as it has proven highly effective for ALDi estimation (Sakr et al., 2024). Its [CLS] representation feeds two parallel heads: a regression head that maps it to a scalar in $[0, 1]$, and a classification head that maps it into K discretized ALDi bins. Only the regression head is used at inference; the classification head shapes the shared encoder during training and is then discarded.

3.2 Binning Strategy

A key challenge in regression tasks is ensuring balanced target distributions across folds during cross-validation. Since standard stratification cannot be applied to continuous variables, we employed a binning strategy serving two purposes:

- **Enabling Stratified Splitting:** Continuous ALDi scores were discretized into K bins, allowing stratified 5-fold cross-validation that preserves the score distribution across splits.
- **Creating an Auxiliary Task:** These same bins act as target labels for the auxiliary classification head, providing a structured coarse-grained learning signal.

We experimented with both equal-width (cut)³ and quantile-based (qcut) binning, ultimately selecting equal-width discretization with $K = 7$ bins. This choice yielded stable training. Figure 1 illustrates how the 7-class setup effectively balances between overly coarse 4-class and overly sparse 10-class discretizations.

3.3 Joint Loss Function

Our model is trained by minimizing a weighted combination of regression and classification losses:

$$L = \alpha_{\text{reg}} \cdot L_{\text{reg}} + (1 - \alpha_{\text{reg}}) \cdot L_{\text{cls}},$$

²<https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter>

³<https://pandas.pydata.org/docs/reference/api/pandas.cut.html>

where α_{reg} balances the two tasks.

The regression loss L_{reg} is the Mean Squared Error (MSE):

$$L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

For the auxiliary classification head, we employ Focal Loss (Lin et al., 2020):

$$L_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N (1 - \hat{p}_{i,c})^\gamma \log(\hat{p}_{i,c}),$$

where γ focuses training on harder examples. Our best-performing model uses $\gamma = 3$ and $\alpha_{\text{reg}} = 0.73$, emphasizing regression while retaining benefits from the classification signal.

3.4 Dataset and Preprocessing

We used the official dataset from the NADI 2024 Shared Task 2: ALDi Estimation (Abdul-Mageed et al., 2024), derived from AOC-ALDi (Keleg et al., 2023), itself built on the Arabic Online Commentary (AOC) corpus (Zaidan and Callison-Burch, 2011). The dataset consists of 102,886 training samples, 107 development samples, and 857 blind test sentences, each annotated with a continuous dialectness score in the range $[0, 1]$. Note that a score of 0 does not reflect a natural “pure MSA” variety—MSA is a formal written standard with no native speakers—but is an annotation construct in the ALDi framework (Keleg et al., 2023): it marks sentences with no dialectal features detectable by human annotators. Text was tokenized using the AraBERTv2-Twitter tokenizer, and sequences were padded or truncated to a maximum of 256 tokens.

3.5 Training and Evaluation Protocol

We employed 5-fold stratified cross-validation using the 7-bin discretization described in Section 3.2. The full multi-task model—the shared AraBERTv2-Twitter encoder trained jointly with both the regression and classification heads—was trained for 5 epochs with the AdamW optimizer (learning rate 4×10^{-6} , weight decay 0.01, batch size 32), a polynomial learning rate scheduler, gradient clipping (max norm 1.0), and dropout regularization.

We tuned dropout rates ($\{0.1, 0.2, 0.25, 0.4\}$), the Focal Loss focusing parameter $\gamma \in \{2, 3, 4\}$, and the regression–classification loss weight

$\alpha_{\text{reg}} \in \{0.5, 0.73, 0.77, 0.88, \dots\}$. We compared binning granularities with $K \in \{4, 7, 10\}$. The best model per fold was chosen based on validation RMSE, since regression is the primary objective and final predictions on the test set were obtained using only the regression head output.

Model performance was evaluated using RMSE. The NADI 2024 blind test set is not released with public ground-truth labels; predictions are scored by submitting them to the official shared-task platform. As the shared task’s evaluation period has concluded, we obtained our blind test result through the platform’s post-evaluation phase, which returns RMSE as the sole metric. Reporting alternative metrics on the blind test set is therefore not possible, and we follow prior work in using RMSE as the benchmark measure.

4 Experimental Results

4.1 Progressive Refinement of the Auxiliary Task

We systematically evaluated the progressive refinement of our MTL framework. The performance at each stage, benchmarked against a pure regression baseline, is detailed in Table 1.

Our baseline model, trained solely on the regression task with MSE loss, achieved a dev RMSE of 0.13746. Our initial MTL (Stage 2), which combined a standard Cross-Entropy classification loss with the MSE regression loss, failed to improve performance and resulted in a higher dev RMSE of 0.13908.

Introducing Focal Loss for the auxiliary task and Huber loss for regression (Stage 3) reduced dev RMSE to 0.13652, indicating that loss shaping is critical for effective multi-task supervision.

Building on this success, we focused on refining the auxiliary task. In Stage 4, increasing the classification granularity from 4 to 7 bins led to a substantial performance gain, lowering the dev RMSE to 0.12857. Further, increasing the Focal Loss focusing parameter to $\gamma = 3$ pushed the performance even further, achieving a dev RMSE of 0.11933.

Finally, in Stage 5, we optimized the synergy between the two tasks. By fine-tuning the regression loss weight to $\alpha_{\text{reg}} = 0.73$, reverting the regression loss to MSE, and adopting a polynomial learning rate scheduler, we achieved our best results. This final model obtained a dev RMSE of **0.10784** and, most importantly, a blind test set RMSE of **0.12754**, demonstrating a signifi-

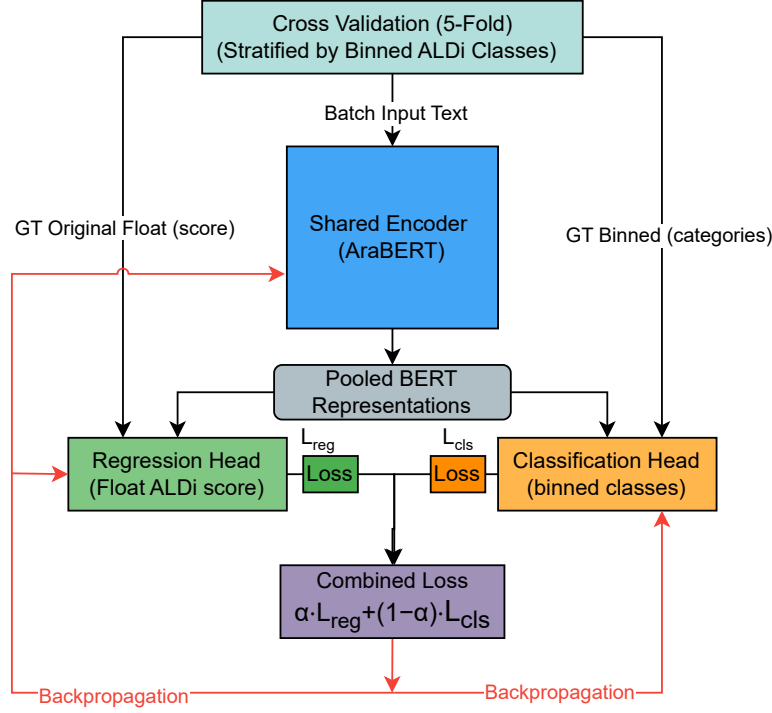


Figure 2: Training flow of the proposed multi-task framework for Arabic Level of Dialectness (ALDi) estimation.

Table 1: Progressive Refinement of an Auxiliary Classification Task for ALDi Regression.

Stage	Focus	Loss (C/R)	Classes	γ	α	Scheduler	Validation RMSE	Blind Test RMSE
1. Baseline	Pure Regression (no auxiliary head)	-/MSE	4	N/A	1.0	cosine	0.13746	N/A
2. First Attempt	Naive auxiliary classification fails	CE/MSE	4	N/A	0.77	linear	0.13908	N/A
3. First Improvement	Focal Loss for the auxiliary head	Focal/Huber	4	2	0.77	cosine	0.13652	N/A
4. Refinement	Refining the auxiliary task							
	a) Increased Granularity	Focal/Huber	7	2	0.77	cosine	0.12857	N/A
	b) Focused on Hard Examples	Focal/Huber	7	3	0.77	cosine	0.11933	0.14037
5. Final Synergy	Optimizing the Training Balance	Focal/MSE	7	3	0.73	polynomial	0.10784	0.12754

Abbreviations: C/R = Classification/Regression; CE = Cross-Entropy; γ = Focal Loss Gamma; α = Regression Loss Weight.

cant improvement over the baseline. The gap between validation (0.10784) and blind test (0.12754) RMSE is expected: the development set contains only 107 sentences and, as noted in prior work on the same task (Sakr et al., 2024), is not fully representative of the 857-sentence blind test distribution. This discrepancy affects all systems on this benchmark; for comparison, AlexUNLP-STM (Sakr et al., 2024) reported a substantially larger gap (development RMSE 0.1060 versus blind test 0.1406), while our blind test result remains the best reported on the benchmark.

4.2 Ablation Beyond Tuning

To address concerns regarding whether improvements stem from hyper-parameter tuning or the auxiliary task architecture, we perform a systematic ablation study starting from our final optimized system. In this experiment, all hyper-parameters—including the polynomial scheduler, learning rate (4×10^{-6}), and regression weight ($\alpha_{\text{reg}} = 0.73$)—are held constant. We only vary the presence and type of the auxiliary classification module, and we repeat the ablation on a second Arabic encoder, MARBERTv2⁴, to test whether any effect is tied to a single backbone. The right-

⁴<https://huggingface.co/UBC-NLP/MARBERTv2>

most column of Table 2 extends the same ablation to English and is discussed in Section 4.4.

Auxiliary supervision improves regression performance under both Arabic encoders. On AraBERTv2-Twitter, even a naive auxiliary head trained with standard Cross-Entropy reduces validation RMSE from 0.1444 to 0.1154, and adding Focal Loss brings it to 0.1078. On MARBERTv2, Focal Loss and Cross-Entropy perform comparably (0.1269 vs. 0.1268), but the MTL framework still delivers a 9.6% gain over pure regression. This suggests that a different encoder can mitigate some label imbalance natively, while the auxiliary classification task remains the primary driver of the improvement. The auxiliary task itself helps under every encoder we test, while the size of the additional margin from Focal Loss varies.

One caveat applies to the pure-regression row. Its 0.1444 is obtained under the same optimization settings as the Final MTL system (polynomial scheduler, identical learning rate, dropout, and training regime), and therefore differs from the Stage 1 baseline in Table 1, which used a cosine scheduler and different optimization choices. The purpose here is not to re-tune pure regression, but to isolate the architectural contribution of the auxiliary task under identical training conditions.

4.3 Comparison with Prior Work

Table 3 summarizes our results against top-performing systems from NADI 2024 Subtask 2. The best previous system, ASOS, achieved a blind test RMSE of 0.1403 using MARBERT-based regressors with translation-driven data augmentation. The AlexUNLP-STM system reported a blind test RMSE of 0.1406, tackling the skewed target distribution with data-level strategies, weighted random sampling and undersampling of the majority MSA class by roughly 43,000 sentences, on top of a contrastive-learning regression ensemble. Its final system was built on the same AraBERTv2-Twitter-Base encoder we use, which makes the comparison direct and controlled. Our approach applies no data-level manipulation and instead handles the skew through auxiliary task design, reducing the error to 0.1275, a relative improvement of more than 9% over the strongest prior result. As a reference point, the Sentence-ALDi model (Kellegher et al., 2023) scored 0.2178 on this blind test set as reported in the NADI 2024 overview (Abdul-Mageed et al., 2024), which our approach improves upon by roughly 41% relative.

4.4 Generalization Beyond ALDi

To test whether our findings extend beyond Arabic dialectness modeling, we applied the framework to the Jigsaw toxicity dataset⁵, a completely different language (English) and domain (toxicity detection). All training settings were kept identical to our final system, except the backbone, which was replaced with RoBERTa-base⁶ since the task language changed from Arabic to English. The dataset contains 712,388 training and 77,298 test samples, where each label is a continuous toxicity score in $[0, 1]$. The dataset is also heavily skewed: 72.2% of training samples and 72.5% of test samples carry an exact score of zero, compared to 52.3% in NADI 2024. The rightmost column of Table 2 reports the results. The performance hierarchy observed on AraBERTv2-Twitter, Focal Loss < Cross-Entropy < Pure Regression in RMSE, is preserved on a completely different language and domain. This is consistent with our core hypothesis that discretization acts as a regularizer preventing the model from collapsing toward the majority mean, and suggests the finding is not specific to Arabic dialectness modeling.

4.5 Inference Cost

A practical concern with multi-task models is deployment overhead. Because the classification head is dropped at inference and only the regression head is used, our final system has the same inference cost as a plain regression model. Table 4 confirms this: latency and memory are virtually identical for the two, so the auxiliary task adds complexity only during training and introduces no overhead at deployment.

5 Model Behavior and Error Analysis

To understand the behavior of our multi-task framework, we perform a dual-stage error analysis: a quantitative assessment of regression bias and a qualitative review of specific linguistic failures.

5.1 Quantitative Analysis

We analyze model behavior across the ALDi spectrum using the diagnostic plots shown in Figure 3, with the goal of identifying systematic biases introduced by skewed supervision.

⁵<https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data>

⁶<https://huggingface.co/FacebookAI/roberta-base>

Model Configuration	Arabic — NADI 2024		English — Jigsaw
	AraBERTv2-Tw. (final system)	MARBERTv2 (robustness)	RoBERTa-base (generalization)
Final MTL System (Focal Loss)	0.1078	0.1269	0.1219
– replace Focal Loss with Cross-Entropy	0.1154	0.1268	0.1231
– remove Auxiliary Task (Pure Regression)	0.1444	0.1404	0.1254

Table 2: Controlled ablation; bold marks our proposed system. All optimization parameters are fixed to the Final System’s settings; only the auxiliary module varies. Arabic columns report validation RMSE, the English column test RMSE; the two blocks are separate experiments and are not comparable.

System	Blind Test RMSE
Ours (MTL, 7 bins, polynomial)	0.1275
ASOS (Nacar et al., 2024)	0.1403
AlexUNLP-STM (Sakr et al., 2024)	0.1406
Sentence-ALDi (Keleg et al., 2023)	0.2178

Table 3: Comparison of blind test RMSE with prior systems on NADI 2024 Subtask 2.

Model	GPU	CPU	VRAM
Pure Regression	3.780	293.691	0.865
Final MTL System	3.686	294.224	0.864

Table 4: Inference cost per sentence. GPU and CPU latency in ms/sentence, VRAM in GB. Dropping the classification head at inference leaves deployment cost unchanged.

First, the **RMSE per Region** (Figure 3a) reveals that the model is most precise in the MSA (0.0 – 0.1) and mid-dialect (0.5 – 0.8) ranges. However, error significantly increases in the extreme dialect range (0.9–1.0), likely due to the sparsity of “pure” dialectal examples in the training set.

Second, the **Residual Plot** (Figure 3b) exposes a clear “regression to the mean” bias. The downward slope indicates that the model systematically over-predicts low ALDi scores and under-predicts high ALDi scores. This suggests the model is conservative, pulling extreme values toward the global average.

Finally, the **Error Boxplot** (Figure 3c) confirms that while the variance of error is tight for MSA and Low Dialect categories, the “High Dialect” category (Bin 6/7) exhibits the widest spread and highest median error, identifying it as the primary area for future data augmentation. All diagnostic analyses in this section are conducted on the official NADI 2024 development set.

5.2 Qualitative Analysis

The examples in Table 5 illustrate systematic differences between the multi-task model (PRD-MTL) and the pure regression baseline (PRD-Reg). These differences help explain why the MTL framework achieves superior stability. We categorize these into three linguistic performance areas based on dialect intensity.

Area 1: The Conservative Wall (Examples 1–2)

In the extreme dialect range ($GT > 0.85$), both models suffer from the “regression to the mean” bias identified in Figure 3b. For Example 1 (transliterated English), the Pure Regression model actually predicts a higher score (0.854) than the MTL model (0.689). However, this baseline behavior is often erratic. While the MTL model remains somewhat “conservative” at the extremes, it avoids the high-variance fluctuations seen in the pure regression baseline across the rest of the dataset.

Area 2: MTL as an Anchor (Examples 3–5)

This range highlights the primary strength of auxiliary classification. In Examples 3, 4, and 5, the Pure Regression model significantly over-predicts dialect intensity (e.g., 0.916 vs. GT 0.666 in Example 4). By contrast, the MTL model “anchors” the prediction within the correct discretized bin. The auxiliary signal prevents the regression head from over-reacting to specific dialectal tokens, resulting in predictions that are significantly closer to the ground truth.

Area 3: Precision and Calibration (Examples 6–8)

In the moderate dialect range (0.5 – 0.7), the MTL model achieves near-perfect calibration. While the Pure Regression baseline remains consistently “noisy”, the MTL predictions in Examples 7 and 8 are identical or nearly identical to the ground truth. This demonstrates that the 7-bin auxiliary task provides a stable coordinate system that the shared en-

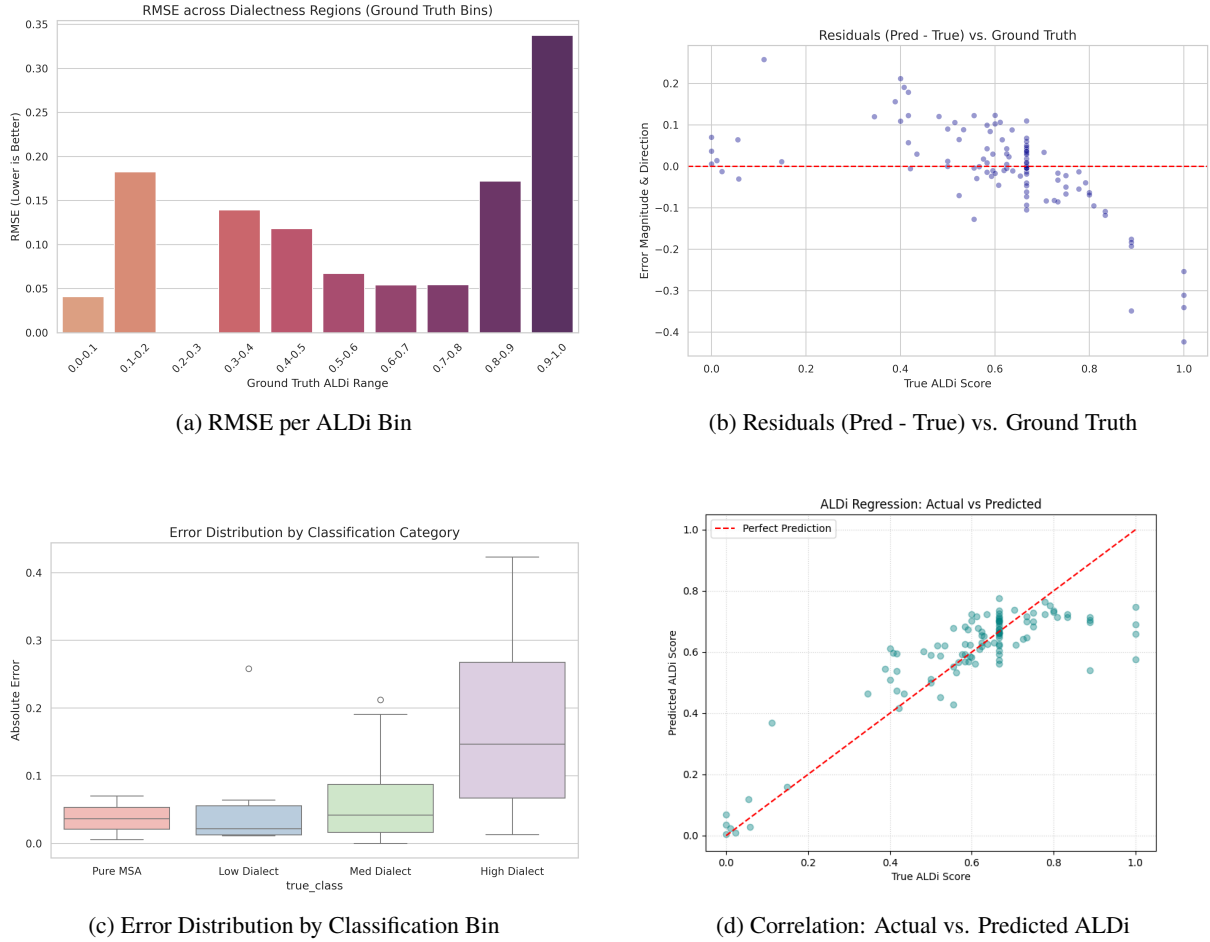


Figure 3: Quantitative error analysis of the Stage 5 MTL model.

coder uses to refine its continuous outputs.

6 Discussion

Why naive auxiliary classification fails. The failure of Stage 2 suggests that a simple classification objective introduces conflicting gradients into the shared encoder: standard Cross-Entropy rewards the model for separating easily classifiable examples, which provides little value for the fine-grained predictions the regression head must make. Focal Loss reverses this. By down-weighting the loss from easy-to-classify examples, it forces the encoder to concentrate on the ambiguous cases lying near bin boundaries — exactly the region where the continuous target is hardest to place.

Sensitivity to the two design choices. Binning granularity has an empirically optimal value: increasing the number of bins from 4 to 7 lowers validation RMSE from 0.13652 to 0.12857, while moving to 10 bins introduces data sparsity in the minority classes (Figure 1), which fixes $K = 7$ as the balance between signal richness and class

density. The focal-loss focusing parameter is more forgiving: both $\gamma = 2$ and $\gamma = 3$ yield substantial gains over the baseline, and the further improvement from $\gamma = 2$ (0.12857) to $\gamma = 3$ (0.11933) reflects a stable optimization trend rather than an artefact of hyperparameter tuning. The loss weight $\alpha_{\text{reg}} = 0.73$ completes the picture: regression remains the primary objective, but 27% of the learning signal still comes from the auxiliary head.

What the auxiliary signal provides. Table 5 shows what this buys in practice. Pure regression, even under an optimized schedule, produces high-variance predictions that drift away from the ground truth when it meets strong dialectal markers; the auxiliary head anchors the prediction inside the correct discretized bin and damps that drift. This is the mechanism behind the error reduction in Table 2, and it holds on both Arabic encoders and on English.

No	Text	PRD-MTL	PRD-Reg	GT
1	I will say let's face-to-face after days. اقولها لتس فيس تو فيس اقدر ديز	0.689	0.854	1.0
2	You disappointed us; may God disgrace you all, from Perez to the youngest player. #Fail #Hala_Madrid فزرونا الله يفضح عرضكم من بيريز لاصغر لاعب #خزوق #هالا-مدريد	0.540	0.555	0.888
3	Why did the topic suddenly turn into cardio? هو الموضوع قالب ع كارديو ليه كده	0.736	0.907	0.800
4	Not you, because I can't stand you — you'd eat my whole bag of chips. بلاش انتي عشان مش طيقاكي اكله باكت البطاطس بتاعي كله	0.776	0.916	0.666
5	When I saw the time, I remembered I had a course I was supposed to drop and forgot. يوم شفت الوقت الساعة تذكرت عندي ماده كنت بسحبها ونسيت	0.648	0.811	0.666
6	The doorman of our college halls probably knows us because of how much we're being tested. ممكن يواب قاعات كليتنا عقد ما بنمتحن	0.667	0.722	0.666
7	The most important thing is that there's a loan on it. اهم شي عليه قرض	0.566	0.614	0.566
8	Happy New Year, my well-mannered neighbor. May God rid you of Majed Al-Harbash. Wishing you health and well-being. كل عام وانت بخير جاري الخلووق والله يفكك من ماجد الحريش وانت بخير وصحه وسلامه	0.500	0.443	0.500

Table 5: Qualitative Comparison: Predictions from the Final MTL model (PRD-MTL) and the corresponding Pure Regression model from the ablation study (PRD-Reg) to isolate the effect of the auxiliary task, against Ground Truth (GT). Bold marks the prediction closer to GT.

7 Conclusion

In this work, we analyzed how auxiliary classification benefits continuous ALDi estimation. Our optimized MTL framework, featuring 7-bin discretization and Focal Loss, achieved a blind test RMSE of 0.1275—a new state-of-the-art for the NADI 2024 benchmark and a 9% relative improvement over prior work.

Our comparison between MTL and Pure Regression models reveals that the auxiliary task serves as a critical calibration tool, significantly reducing prediction noise in the moderate-to-high dialect range. While a “conservative wall” persists at the extreme ends of the spectrum, the design transfers across encoders and to English, suggesting it is a useful starting point for modeling continuous linguistic scales under heavily skewed supervision.

Limitations

While our study provides strong empirical evidence on the ALDi benchmark, several limitations should be noted.

First, our Arabic evidence comes from a single dataset. ALDi is, to the best of our knowledge, the only Arabic resource with sentence-level, human-annotated continuous scores, so cross-dataset validation within Arabic is not currently possible, and the scarcity of Arabic continuous annotations lim-

its systematic evaluation across tasks. Our English experiment on Jigsaw (Section 4.4) shows the framework transfers to one other skewed regression task, but a single out-of-language check does not establish that our conclusions hold for regression problems with different label distributions or linguistic properties. Second, both Arabic encoders we study are base-sized masked language models. Larger or instruction-tuned models may interact differently with auxiliary supervision, and even within this family the margin by which Focal Loss improves on Cross-Entropy varies with the encoder. Third, our study focuses exclusively on improving ALDi prediction accuracy. We do not evaluate whether improved dialectness estimation translates into downstream gains for tasks such as sentiment analysis or machine translation. Arabic punctuation restoration is one concrete candidate: such systems are trained on MSA-heavy corpora and degrade sharply under register shift (Sakr and Torki, 2023; Al Wazrah et al., 2025), so a per-sentence dialectness score could flag how far an input lies from that training distribution. We leave these directions to future work.

References

Muhammad Abdul-Mageed, Amr Keleg, Abdel-Rahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar

- Habash. 2024. [NADI 2024: The fifth nuanced Arabic dialect identification shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. [NADI 2022: The third nuanced Arabic dialect identification shared task](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Asma Ali Al Wazrah, Afrah Altamimi, Hawra Al-jasim, Waad Alshammari, Rawan Al-Matham, Omar Elnashar, Mohamed Amin, and Abdulrahman AlOsaimy. 2025. Evaluation of large language models on arabic punctuation prediction. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 144–154.
- Axel Berg, Magnus Oskarsson, and Mark O’Connor. 2021. [Deep ordinal regression with label diversity](#). In *2020 25th International Conference on Pattern Recognition (ICPR)*, page 2740–2747. IEEE.
- Paula Branco, Luis Torgo, and Rita P Ribeiro. 2019. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99.
- Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.
- Shijie Chen, Yu Zhang, and Qiang Yang. 2024. [Multi-task learning in natural language processing: An overview](#). *Preprint*, arXiv:2109.09138.
- Arwa Fadel, Mostafa Saleh, Reda Salama, and Osama Abulnaja. 2024. [Mtl-arabert: An enhanced multi-task learning model for arabic aspect-based sentiment analysis](#). *Computers*, 13(4).
- Pasquale Foggia, Antonio Greco, Alessia Saggese, and Mario Vento. 2023. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Engineering Applications of Artificial Intelligence*, 118:105651.
- Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.
- Junguang Jiang, Baixu Chen, Junwei Pan, Ximei Wang, Dapeng Liu, Jie Jiang, and Mingsheng Long. 2023. [Forkmerge: Mitigating negative transfer in auxiliary-task learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 30367–30389. Curran Associates, Inc.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. [Focal loss for dense object detection](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. [Multi-task deep neural networks for natural language understanding](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.
- Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Tamar Solorio. 2016. [Overview for the second shared task on language identification in code-switched data](#). In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.
- Omer Nacar, Serry Sibae, Abdullah Alharbi, Lahouari Ghouti, and Anis Koubaa. 2024. [ASOS at NADI 2024 shared task: Bridging dialectness estimation and MSA machine translation for Arabic language enhancement](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 748–753, Bangkok, Thailand. Association for Computational Linguistics.

- Silvia L Pinteá, Yancong Lin, Jouke Dijkstra, and Jan C van Gemert. 2023. A step towards understanding why classification helps regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19972–19981.
- Ruizhi Pu, Gezheng Xu, Ruiyi Fang, Bing-Kun Bao, Charles Ling, and Boyu Wang. 2025. Leveraging group classification with descending soft labeling for deep imbalanced regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19978–19985.
- Noah C Puetz, Jens U Brandt, Marc Hilbert, Elena Raponi, Thomas Bäck, and Thomas Bartz-Beielstein. 2026. Deconstructing deep imbalanced regression: a comprehensive review and experimental evaluation. *Artificial Intelligence Review*.
- A. Sakr and M. Torki. 2023. [Arapunc: Arabic punctuation restoration using transformers](#). In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.
- Abdelrahman Sakr, Marwan Torki, and Nagwa El-Makky. 2024. [AlexUNLP-STM at NADI 2024 shared task: Quantifying the Arabic dialect spectrum with contrastive learning, weighted sampling, and BERT-based regression ensemble](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 735–741, Bangkok, Thailand. Association for Computational Linguistics.
- Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. [Fine-grained Arabic dialect identification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. [Overview for the first shared task on language identification in code-switched data](#). In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.
- Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. 2021. Delving into deep imbalanced regression. In *International conference on machine learning*, pages 11842–11851. PMLR.
- Omar F. Zaidan and Chris Callison-Burch. 2011. [The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.
- Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. 2023. [Improving deep regression with ordinal entropy](#). *Preprint*, arXiv:2301.08915.