KvochurHegel at AbjadMed: Combining LDAM Loss and Adversarial Training for Arabic Medical Question-Answer Classification

Minh-Hoang Le
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
24520542@gm.uit.edu.vn

Abstract

This paper describes our team's submission to AbjadMed at AbjadNLP 2026. The task involves classifying Arabic medical question-answer pairs into 82 categories, characterized by a long-tail distribution and significant semantic overlap. While domain-specific Arabic models exist, they are primarily optimized for Named Entity Recognition or span-extraction tasks rather than high-cardinality sequence classification. Consequently, our system adopts a robust optimization approach using a general-purpose encoder. We utilize **ARBERTv2** as the backbone, employing **Label-Distribution-Aware Margin (LDAM) loss** to mitigate class imbalance and **Fast Gradient Method (FGM)** adversarial training to enhance generalization boundaries. Our approach achieves a Macro-F1 score of 0.4028 on the private test set, demonstrating that advanced optimization techniques can yield competitive performance on specialized taxonomies without requiring domain-specific pre-training.

1 Introduction

Medical question-answer classification systems play a crucial role in healthcare information retrieval, enabling automated triage and clinical decision support. While English medical NLP benefits from robust biomedical encoders such as ClinicalBERT and BioBERT, Arabic medical NLP remains under-resourced, with most existing models focusing on specific tasks rather than providing general-purpose representations suitable for diverse medical taxonomies.

The AbjadMed (Gupta et al., 2026) addresses this gap by requiring classification of Arabic medical question-answer pairs into 82 categories. The dataset presents two key challenges: (1) severe class imbalance with a long-tail distribution, where some categories have fewer than 10 training examples while others have hundreds; and (2) semantic ambiguity between related categories, where decision boundaries are inherently unclear. For instance, distinguishing *Dental Diseases* (Class 13) from *Dentistry* (Class 15) requires nuanced understanding beyond simple keyword matching.

Our approach combines ARBERTv2, a general-purpose Arabic encoder, with techniques specifically designed to address these challenges. To handle class imbalance, we employ Label-Distribution-Aware Margin (LDAM) loss, which dynamically adjusts decision margins based on class frequency. To improve robustness against ambiguous labels and noisy training signals, we integrate Fast Gradient Method (FGM) adversarial training. Additionally, we apply manual re-weighting to highly confusable category pairs identified through validation analysis.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 describes the shared task and dataset characteristics, Section 4 details our methodology, Section 5 presents results and error analysis, and Section 6 discusses limitations and concludes.

2 Background

Task and Data: The AbjadMed (Gupta et al., 2026) benchmarks Arabic medical question-answer classification. The dataset contains 27,951 training samples and a blind test set of 18,634 samples (split equally into public and private evaluation sets). It presents two primary challenges: (1) severe class imbalance with a long-tail distribution where 39 "head" classes have 600 samples each while minority classes like *Biochemistry* have as few as 7 (imbalance ratio ≈86:1), and (2) semantic ambiguity between overlapping categories such as *Dental Diseases* (13), *Dental Health* (14), and *Dentistry* (15), which share significant lexical overlap.

Related Work: Generic Arabic models like ARBERTv2 (Abdul-Mageed et al., 2021) lack spe-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 120–123
March 28, 2026. ©2026 Association for Computational Linguistics

cialized medical vocabulary, and domain-specific pre-training remains limited due to scarce annotated corpora. Common methods for handling class imbalance include re-sampling, cost-sensitive re-weighting (Cui et al., 2019), and loss modification techniques like Focal Loss (Lin et al., 2017). However, margin-based losses like LDAM (Cao et al., 2019) have shown superior generalization for deep encoders by enforcing larger decision boundaries for minority classes. Additionally, adversarial training (FGM), adapted for NLP by Miyato et al. (2017), acts as a regularizer to smooth decision boundaries and improve robustness against label noise.

3 Methodology

We address the challenges of class imbalance and semantic ambiguity through a combination of architectural choices and targeted optimization techniques. Our system uses ARBERTv2 (Abdul-Mageed et al., 2021; Elmadany et al., 2022) as the base encoder with custom head-tail truncation, optimized using Layer-wise Learning Rate Decay (LLRD) (Howard and Ruder, 2018), adversarial training (FGM) (Miyato et al., 2017), and a modified Label-Distribution-Aware Margin (LDAM) loss (Cao et al., 2019).

3.1 Model Architecture

We utilize ARBERTv2 (Abdul-Mageed et al., 2021; Elmadany et al., 2022), a BERT-Base model pre-trained on 243GB of Modern Standard Arabic (MSA) text. We extract the `[CLS]` token embedding ($d = 768$) and pass it through a linear classification head to compute probabilities over the 82 categories.

3.2 Head-Tail Truncation

Standard truncation (keeping only the first 512 tokens) is not ideal for medical QA, as the doctor's final recommendation often appears at the end. We construct inputs as `[CLS] Question [SEP] Answer [SEP]`. To preserve both the patient's complaint and the doctor's conclusion within the 512-token limit, we reserve at least 50 tokens for the answer. If the answer exceeds the remaining space, we retain the first and last 50% of the available answer tokens.

3.3 Optimization Strategy

We employ two regularization techniques to stabilize training on the noisy dataset.

Layer-wise Learning Rate Decay (LLRD): Following Howard and Ruder (2018), we assign a higher learning rate (η_{head}) to the classification head and decay the rate for lower layers layer-by-layer. The learning rate for layer l is defined as:

$$\eta_l = \eta_{head} \cdot \xi^{L-l} \tag{1}$$

where $\xi = 0.9$ is the decay factor and $L = 12$ is the number of layers.

Adversarial Training (FGM): To improve robustness, we apply the Fast Gradient Method adapted for text (Miyato et al., 2017). We compute the perturbation r_{adv} as:

$$r_{adv} = \epsilon \cdot \frac{g}{||g||_2}, \quad g = \nabla_x \mathcal{L}(f(x;\theta), y) \tag{2}$$

We set $\epsilon = 0.2$. This forces the model to learn smooth decision boundaries in the embedding space.

3.4 Class Imbalance Handling

We address the long-tail distribution using **LDAM Loss** (Cao et al., 2019). LDAM enforces a class-dependent margin Δ_y that is inversely related to the class frequency n_y:

$$\mathcal{L}_{\text{LDAM}} = -\log \frac{e^{z_y - \Delta_y}}{e^{z_y - \Delta_y} + \sum_{j \neq y} e^{z_j}} \tag{3}$$

where $\Delta_y = C/n_y^{1/4}$. We use the default $C = 1.0$ and apply Deferred Re-Weighting (DRW), training with standard Cross-Entropy for 5 epochs before switching to LDAM.

Targeted Margin Boosting: Validation analysis revealed high confusion among semantically overlapping categories in the Dental, Pediatric, and Psychiatric domains. These "head" classes (600 samples each) suffered from high inter-class ambiguity. To enforce sharper decision boundaries, we artificially reduced their frequency counts ($n_y \to n_y/10$) *exclusively* for the margin calculation (Δ_y). Crucially, we decoupled this from the DRW schedule, which continued to use the original frequencies. This ensures that while the decision boundaries for ambiguous classes are pushed outward, the overall loss weighting remains stable.

3.5 Experimental Setup

We trained our model on a single NVIDIA A100 (40GB) GPU using mixed-precision (BF16). The

model was trained for 8,000 steps (approx. 20 epochs) with a batch size of 64. We used the AdamW optimizer with a linear warmup (10% of steps) and set the random seed to 42 for reproducibility.

4 Results and Analysis

4.1 Main Results

Our model achieved a **Macro-F1 of 0.4028** on the private test set (Rank 16/42), surpassing our public score of 0.3968. This improvement on the private set suggests that our regularization strategies (FGM and LDAM) successfully prevented overfitting.

4.2 Per-Class Performance

To understand our model's behavior across the imbalanced distribution, we analyzed per-class F1 scores on our local validation set (80/20 split). Table 1 highlights representative categories.

High-frequency classes with distinct medical terminology achieved strong performance (e.g., *Addiction*: 0.86, *Diabetes*: 0.84). However, majority classes with high semantic overlap remained challenging despite targeted margin boosting: *Dental Diseases* (F1: 0.27) and *Dentistry* (F1: 0.48) performed significantly worse than unambiguous classes of similar size. Finally, severe minority classes (<10 samples) failed completely, with categories like *Biochemistry* and *Vascular Surgery* achieving F1 scores of 0.00, indicating that loss re-weighting alone cannot overcome severe data scarcity.

Category	Count	Val F1	Observation
Distinct Majority Classes			
Addiction	600	0.86	Distinct symptoms
Diabetes	600	0.84	Clear terminology
Ambiguous Majority Classes			
Child Health	600	0.50	Confused w/ Pediatric
Dental Diseases	600	0.27	Confused w/ Dentistry
Extreme Minority Classes			
Biochemistry	7	0.00	Data insufficient
Vascular Surg.	7	0.00	Data insufficient

Table 1: Validation F1 scores for selected categories representing distinct performance clusters. Count refers to training samples.

5 Discussion and Conclusion

Our results on the AbjadMed (Gupta et al., 2026) demonstrate that general-purpose Arabic encoders can achieve competitive performance (Rank 16/42) on specialized medical taxonomies when paired with robust optimization strategies. By combining LDAM loss (Cao et al., 2019) with adversarial training, we successfully recovered performance for "middle-tail" classes without relying on domain-specific pre-training.

However, our analysis defines clear boundaries for this optimization-centric approach:

- **Extreme-Tail Limit:** Loss re-weighting hits a hard limit with classes having fewer than 10 samples, which remained unlearnable in our experiments.
- **Semantic Ambiguity:** Near-synonymous categories (e.g., *Dental Diseases* vs. *Dentistry*) persist despite margin enforcement, indicating that lexical overlap outweighs margin separation.
- **Scalability:** Our reliance on manual heuristic adjustments for these ambiguous pairs is effective but not automatically scalable to new datasets.

In summary, while our system establishes a competitive standard for the task, the fundamental challenges of extreme data scarcity and semantic ambiguity in Arabic medical NLP remain significant hurdles that likely require data-centric rather than model-centric solutions. Furthermore, given the critical nature of healthcare, such systems should function as decision-support tools subject to human verification rather than autonomous diagnostic agents.

Acknowledgments

We thank the organizers of the AbjadMed Shared Task at AbjadNLP 2026 for providing the dataset and facilitating the benchmark. We also thank the anonymous reviewers for their valuable feedback which helped improve the quality of this work.

References

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.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. *Preprint*, arXiv:1901.05555.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2022. Orca: A challenging benchmark for arabic language understanding. *arXiv preprint arXiv:2212.10758*.

Pranav Gupta, Niranjan Kumar M, Balaji Nagarajan, Imed Zitouni, and Mo El-Haj. 2026. Abjadmed: Arabic medical text classification at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.