Sujith Kanakkassery at AbjadMed: Imbalance-Aware Transformer Fine-tuning for Arabic Medical Text Classification

Sujith Kanakkassery
sj.kanakkassery@gmail.com

Abstract

This paper describes our system submitted to the AbjadMed 2026 shared task at AbjadNLP. The task focuses on the multi-class classification of Arabic medical texts under severe class imbalance. Our approach fine-tunes a pre-trained Arabic Transformer model and incorporates several imbalance-aware strategies, including data cleaning, class-weighted loss, and label smoothing. Through ablation experiments, we observe consistent improvements over a baseline system, demonstrating the effectiveness of these techniques in improving performance on underrepresented medical categories. Finally, our error analysis highlights persistent challenges related to label sparsity and semantic overlap among medical classes.

1 Introduction

Fine-grained classification of Arabic medical text remains a significant challenge due to semantic overlap between categories, the rarity of specific conditions, and the high variation in question–answer length. These factors often render standard fine-tuning unreliable under macro-averaged metrics. The AbjadMed 2026 shared task (Gupta et al., 2026) addresses this by framing medical classification as a large-scale, highly imbalanced multi-class problem.

Recent work has explored the medical reasoning of Large Language Models (LLMs) in Arabic healthcare (AlDahoul and Zaki, 2025) and model merging to bridge dialectal gaps in clinical settings (Ibrahim et al., 2025). This progress builds on robust pretrained models like CAMeLBERT (Inoue et al., 2021) and MARBERT (Abdul-Mageed et al., 2021). While these models achieved success in tasks like AraHealthQA 2025 (Alhuzali et al., 2025; Zaytoon et al., 2025), those settings used relatively coarse-grained label spaces (e.g., 7 question categories). In contrast, AbjadMed 2026 requires classification across 82 fine-grained medical labels under severe class imbalance, presenting a distinct and more complex challenge.

In this work, we address this label scarcity through a specialized training pipeline incorporating class-weighted loss and label smoothing. Our approach emphasizes robust performance on the long-tail distribution of medical categories, evaluated via macro-averaged F1 score.

2 Task Description

To address the challenges of semantic overlap and label scarcity, the AbjadMed 2026 shared task introduces a large-scale, fine-grained classification benchmark. This task, situated within the main medical track of the AbjadNLP 2026 workshop, focuses on the multi-class classification of Arabic medical narratives.

Each instance in the dataset consists of a question–answer pair provided as a single concatenated text field. These instances are annotated with one of 82 predefined medical categories. While the labels are provided as integer identifiers, the category names—originally in Arabic—were translated into English using a large language model to facilitate analysis and interpretation.

A defining characteristic of this dataset is its highly imbalanced label distribution, which features a small number of high-frequency categories and a significant "long-tail" of low-frequency ones. Consequently, systems are evaluated using the macro-averaged F1 score. This metric ensures that performance on underrepresented medical categories is prioritized, reflecting the real-world complexity of diagnosing rare conditions in a clinical setting.

3 System Overview

3.1 Model Architecture

Our system is built on a pretrained Arabic Transformer encoder with 12 Transformer layers and a

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 408–412
March 28, 2026. ©2026 Association for Computational Linguistics

hidden size of 768. We use the pooled representation produced by the encoder, corresponding to the `[CLS]` token, as a sequence-level embedding. This representation is passed to a lightweight classification head consisting of a two-layer feedforward network with ReLU activation and dropout, followed by a linear projection to the 82 medical categories.

3.2 Preprocessing

We applied lightweight preprocessing to reduce noise while preserving domain-relevant information. Specifically, non-informative conversational phrases such as greetings and boilerplate expressions are removed using a rule-based filtering step. No stemming or normalization is applied in order to retain medically relevant surface forms. The cleaned text is tokenized using the model's native tokenizer, with inputs truncated or padded to a maximum length of 256 tokens.

3.3 Imbalance-Aware Training

To mitigate class imbalance, we trained the model using cross-entropy loss with class weights computed as the inverse frequency of labels in the training set.

We calculated class weights W_c as:

$$W_c = \frac{N}{C \times n_c}$$

where N is the total samples, C is the number of classes, and n_c is the count of samples in class c. Additionally, label smoothing was applied to reduce overconfidence and improve generalization, particularly for rare classes.

4 Experimental Setup

4.1 Dataset

We conducted our experiments on the dataset[1] released by AbjadMed Organizers. The dataset consists of Arabic medical question-answer pairs categorized into 82 distinct medical classes. The official data distribution is as follows:

- **Training Set**: The provided training data contains 27,951 instances. For internal development, we employed a stratified 90/10 split, resulting in 25,156 samples for training and 2,795 samples for local validation.
- **Test Set**: The final evaluation was conducted on the hidden Kaggle test set, which consists of 18,634 instances.

The dataset is characterized by a "long-tail" distribution, where a few majority classes (e.g., Pediatrics, Dermatology) dominate the sample count, while many specialized categories contain very few instances. This extreme class imbalance, combined with the macro-averaged F1 score evaluation metric, necessitates a model that generalizes well across both high-resource and low-resource labels. Figure 1 shows the top twenty performing classes on the validation set.

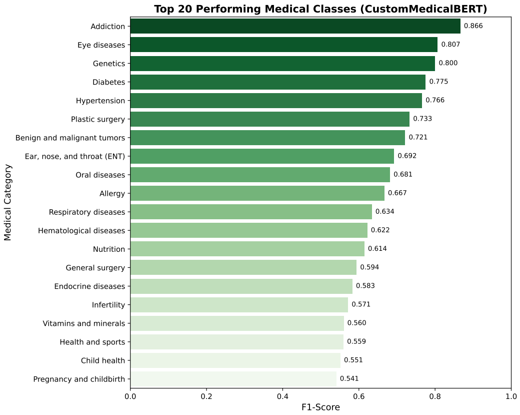

Figure 1: Top 20 medical categories with the highest macro-averaged F1 scores

4.2 Model and Training Setup

All experiments were conducted by fine-tuning pretrained Arabic Transformer models using the Hugging Face `Trainer` framework. We optimized the models using the AdamW optimizer (Loshchilov and Hutter, 2019) with decoupled weight decay.

We initially established a baseline using CAMeLBERT but found that MARBERT consistently achieved superior performance. We attribute this robustness to MARBERT's pretraining on a larger, more diverse, and noisier Arabic corpus, which better captures the linguistic variability relevant to patient-generated medical queries. Consequently, all subsequent experiments utilized MARBERT as the primary backbone, fine-tuned into a configuration we refer to as CustomMedicalBert.

To address the 82-class imbalance and optimize classification performance, we applied several training refinements:

- **Preprocessing:** A rule-based step was implemented to filter non-informative Arabic greet-

[1] `https://balajinaga.github.io/EACL2026-Abjad-NLP-SharedTask/`

ings and boilerplate text.

- **Architecture:** The maximum input length was set to 256 tokens to accommodate longer medical question–answer instances.
- **Regularization:** We employed label smoothing (0.05) and class-weighted loss to prevent majority-class bias.
- **Optimization:** Training utilized a learning rate of 2×10^{-5}, a 0.1 warmup ratio, and a cosine learning rate scheduler.

Based on validation results, the final models were trained for 3 epochs using a batch size of 16 for training and 32 for evaluation. The best checkpoint was selected according to validation macro-averaged F1 score. Table 1 summarizes the final hyperparameter configuration.

Hyperparameter	Value
Base Model	MARBERT (UBC-NLP/MARBERT)
Max Sequence Length	256
Batch Size (per device)	16
Learning Rate	2×10^{-5}
Training Epochs	3
LR Scheduler	Cosine
Warmup Ratio	0.1
Weight Decay	0.01
Label Smoothing	0.05
Seed (Data/Model)	3

Table 1: Final hyperparameter configuration for the CustomMedicalBert Model.

4.3 Computational Environment and Reproducibility

To ensure the transparency and replicability of our findings, we conducted all experiments within a standardized environment provided by the Kaggle platform.

- **Hardware:** Experiments were executed on a single NVIDIA Tesla P100-PCIE (16GB) GPU.
- **Software Stack:** Our pipeline utilized Python 3.12, PyTorch 2.8.0 with CUDA 12.6, and the Hugging Face Transformers (v4.57.1) library.
- **Efficiency:** The final CustomMedicalBert model was trained for 3 epochs, reaching optimal validation performance in approximately 22.38 minutes (1,342.9 seconds).

We fixed a global random seed of 3 to ensure deterministic behavior across all training runs. The complete source code and environment configuration are available in our public repository.[2]

5 Results

5.1 Quantitative Results

Table 2 reports the results of an ablation study conducted on the validation set, evaluating the impact of individual system components on macro-averaged F1 score. We start from a baseline Transformer fine-tuning setup and incrementally introduce imbalance-aware refinements.

System Configuration	Macro-F1
Baseline fine-tuning	0.258
+ Data cleaning	0.265
+ Increased token length	0.293
+ Class-weighted loss	0.320
+ Reduced epochs	0.340

Table 2: Ablation results showing the effect of incremental system refinements on validation Macro-F1.

Each refinement yields consistent performance gains over the baseline. In particular, increasing the input sequence length and incorporating class-weighted loss lead to substantial improvements, highlighting the importance of modeling long medical queries and addressing severe class imbalance. Reducing the number of training epochs further improves generalization by mitigating overfitting, resulting in the best overall Macro-F1 score.

5.2 Error Analysis

Error analysis reveals that most misclassifications occur between semantically related medical categories. Several low-frequency classes exhibit zero F1-score; these correspond exclusively to labels with extremely low support, indicating data sparsity rather than systematic model failure.

Table 3 shows several categories exhibit zero macro-F1 due to extreme label sparsity, with many classes having fewer than five training instances, making reliable learning infeasible.

To analyze systematic errors, we extracted off-diagonal entries from the confusion matrix and ranked misclassified class pairs by frequency. Table 4 reports the most frequent confusions, where

[2] https://github.com/sujik18/EACL-2026-Abjad-NLP

Class ID	Category Name	Support
3	Anatomy	4
6	Biochemistry	1
7	Biology	3
10	Chemistry	1
12	Congenital malformations	1
18	Diagnosis	15
20	Embryology	4
29	Geriatric health	1
34	History of medicine	1
38	In vitro fertilization (IVF)	1
44	Medical services	5
48	Microbiology	3
56	Pathology	3
58	Pediatric surgery	1
60	Physiology	2
64	Preventive medicine	2
69	Ramadan	1
71	Rheumatic diseases	2
75	Toxicology	3
78	Vaccines and immunizations	2
79	Vascular surgery	1

Table 3: Categories with zero macro-F1 score due to extreme label sparsity. Support indicates the number of instances available for each class in the dataset.

Actual (ID)	Predicted (ID)	Count
Pediatric diseases (57)	Child health (11)	37
Sexually transmitted diseases (73)	Sexual health (72)	28
Internal medicine diseases (41)	Gastrointestinal diseases (24)	23
Psychiatric diseases (65)	Mental health (47)	22
Dental diseases (13)	Dental health (14)	21
Women's health (81)	Gynecological diseases (31)	21
Dentistry (15)	Dental health (14)	18
Pregnancy and childbirth (63)	Gynecological diseases (31)	16
Dentistry (15)	Jaw and dental surgery (42)	15
Dental diseases (13)	Dentistry (15)	14

Table 4: Top 10 most frequent misclassifications observed on the validation set. Class IDs are shown in parentheses.

the count denotes the number of evaluation instances in which a ground-truth class was incorrectly predicted as another class. Most frequent misclassifications occur between semantically related medical categories (e.g., pediatric diseases vs. child health, dental diseases vs. dentistry), indicating ambiguity arising from overlapping clinical terminology rather than random model errors.

Metric	Macro Avg	Weighted Avg
Precision	0.3451	0.4782
Recall	0.3673	0.4939
F1-score	0.3408	0.4716

Table 5: Final performance of the proposed system on the validation set.

5.3 Final System Performance

Table 5 reports the final system performance using macro-averaged and weighted-averaged metrics. The gap between macro and weighted scores reflects the severe label imbalance present in the dataset.

6 Conclusion

We presented an imbalance-aware Transformer-based system for Arabic medical text classification in the AbjadNLP Shared Task. Our experiments demonstrate that cost-sensitive learning via class-weighted loss and label smoothing substantially improves macro-F1 performance under extreme class imbalance

As the current system relies on a fixed token length of 256, some longer medical question–answer instances may be truncated. Furthermore, the noise removal process is currently limited to Modern Standard Arabic greetings and may not capture all dialectal variations. Future work will address these limitations by exploring more robust preprocessing strategies, as well as data augmentation and hierarchical classification techniques to better handle rare medical categories.

7 Acknowledgments

We thank the shared task organizers for releasing the dataset that facilitated our experiments, as well as for fostering broader engagement with the task.

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.

Nouar AlDahoul and Yasir Zaki. 2025. NYUAD at AraHealthQA shared task: Benchmarking the medical understanding and reasoning of large language models in Arabic healthcare tasks. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 119–125, Suzhou, China. Association for Computational Linguistics.

Hassan Alhuzali, Walid Al-Eisawi, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Leen Kharouf, Farah E. Shamout, and Nizar Habash. 2025. AraHealthQA 2025: The first shared task on Arabic health question answering. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 107–118, Suzhou, China. Association for Computational Linguistics.

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.

Ahmed Ibrahim, Abdullah Hosseini, Hoda Helmy, Wafa Lakhdhar, and Ahmed Serag. 2025. Bridging dialectal gaps in Arabic medical LLMs through model merging. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 338–346, Suzhou, China. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–101, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR 2019, New Orleans, Louisiana, USA. OpenReview.net.

Mohamed Zaytoon, Ahmed Mahmoud Salem, Ahmed Sakr, and Hossam Elkordi. 2025. AraMinds at AraHealthQA 2025: A retrieval-augmented generation system for fine-grained classification and answer generation of Arabic mental health Q&A. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 198–203, Suzhou, China. Association for Computational Linguistics.