Olga Snissarenko at AbjadMed: Arabic Clinical Text Classification with AraBERT: Results from the AbjadMed Shared Task

Olga Snissarenko
Kazakhstan Branch of Lomonosov Moscow State University
Kazakhstan
snisarenkoola@gmail.com

Abstract

We address Arabic medical text classification into 82 categories under severe class imbalance, where class frequencies range from 7 to 600 samples. Our approach fine-tunes pretrained AraBERT models, comparing configurations with varying text normalization, input length, pooling mechanisms (mean vs. attention), and loss functions. Through systematic experimentation, we find that class-weighted cross-entropy loss provides greater performance gains than architectural enhancements like attention pooling. Our strongest model—AraBERT with mean pooling and balanced class weighting—achieves macro-F1 scores of 0.387 (public) and 0.411 (private), ranking 12th on the shared task leaderboard and highlighting the importance of loss design for imbalanced Arabic clinical text.

1 Introduction

Automatic categorization of medical data is a key component of large-scale healthcare information systems, enabling efficient routing, indexing, and analysis of patient-doctor interactions. In this work, we address the problem of classifying Arabic medical question-answer data written in Modern Standard Arabic into predefined medical categories, following the AbjadMed Shared Task (Gupta et al., 2026).

This task poses several challenges for current modeling techniques. First, the dataset exhibits a highly imbalanced, long-tailed label distribution across 82 categories, with class frequencies ranging from 7 to 600 samples per class. Such imbalance biases standard training objectives toward frequent classes, leading to poor generalization on rare but clinically important categories and motivating the use of class-weighted loss functions. Second, the input texts vary substantially in length and structure, often containing informal explanations, domain-specific medical terminology, and redundant contextual information. Finally, Arabic medical NLP remains relatively underexplored compared to English, due to limited annotated resources and the linguistic complexity of Arabic.

Recent advances in transformer-based language models have significantly improved Arabic language understanding. AraBERT (Antoun et al., 2020) introduced large-scale pretrained transformers tailored for Arabic through extensive pretraining and normalization strategies. Subsequent work demonstrated the effectiveness of fine-tuning AraBERT for a variety of downstream tasks, including classification and information extraction. However, the effectiveness of different modeling and optimization strategies for highly imbalanced, long-form Arabic medical text classification remains insufficiently studied.

In this paper, we systematically compare a set of practical AraBERT-based classification configurations that differ in pooling strategy, input preprocessing, and loss design. Rather than proposing a single novel architecture, our goal is to identify which components are most critical for robust performance under severe class imbalance. We explore Arabic-specific text normalization, attention-based pooling, and class-weighted loss functions. Our results show that addressing class imbalance at the loss level is more impactful than increasing architectural complexity, providing practical guidance for similar low-resource and imbalanced Arabic NLP tasks. The source code for all experiments and model configurations described in this paper is publicly available.[1]

2 Data

The dataset consists of Arabic medical consultation texts structured as question-answer pairs. Each sample contains a single text field that includes both the question, introduced by explicit

[1] https://github.com/Ollasni/Abjad_NLP_Shared_Task_4

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 177–181

Arabic question and answer markers. The texts are written in Modern Standard Arabic and cover a wide range of medical topics, including patient-described symptoms and professional medical advice. The length of the text varies significantly, from brief consultations to long and detailed explanations.

The training set (TRAIN) contains 27,951 labeled samples. Each instance includes a full consultation text, a textual medical category, and an integer label corresponding to one of 82 distinct classes. The development/test set (DEVTEST) contains 18,634 samples with the same text structure but without ground-truth labels. The dataset exhibits strong class imbalance, following a long-tailed distribution, where a small number of classes contain several hundred samples, while many others are sparsely represented.

To ensure reliable evaluation, the training data was split into training and validation subsets using a 90/10 ratio with stratification over class labels. A fixed random seed (42) was used to ensure reproducibility. Prior to tokenization, we applied Arabic-specific text normalization to remove question-answer markers, normalize orthography, remove diacritics, and standardize whitespace. These preprocessing steps were implemented using a custom normalization pipeline tailored for medical and forum-style Arabic text.

3 System

Our system is based on fine-tuning pretrained AraBERT models for multi-class text classification. We experiment with several AraBERT variants, including `aubmindlab/bert-base-arabertv02` and its Twitter-pretrained counterpart, following the standard fine-tuning paradigm for transformer-based models (Antoun et al., 2020).

In all configurations, the AraBERT encoder produces contextualized token representations, which are aggregated into a fixed-length sentence representation using either mean pooling or attention-based pooling. The pooled representation is passed through a dropout layer and a linear classification head mapping to the 82 target classes. Models are trained using the AdamW optimizer with learning rates ranging from 1×10^{-5} to 2×10^{-5}, batch sizes between 8 and 16, and training durations of up to 8 epochs with early stopping based on validation macro-F1 score.

To address the severe class imbalance, we incorporate class-weighted cross-entropy loss, where weights are inversely proportional to class frequencies. All experiments are conducted using a fixed random seed and identical data splits to ensure comparability across configurations.

4 Experimental Setup

4.1 Data Split

The dataset was split into training and validation subsets using a 90/10 ratio with stratification over class labels in order to preserve the original label distribution. A fixed random seed (42) was used across all experiments to ensure reproducibility.

4.2 Evaluation Metric

Model performance was evaluated using the macro-averaged F1 score, which is suitable for imbalanced classification settings as it assigns equal importance to all classes.

5 Models

We conducted a series of experiments based on pretrained AraBERT models, progressively increasing model complexity and incorporating techniques to address class imbalance and representation quality.

5.1 Baseline 1: Single-task AraBERT Classifier

Tokenizer and pretrained model. We used the pretrained AraBERT model `aubmindlab/bert-base-arabertv02-twitter`. Tokenization was performed using the corresponding `AutoTokenizer`, with truncation and padding applied to a maximum sequence length of 128 tokens.

Model architecture. The model consists of a pretrained AraBERT encoder followed by mean pooling over the last hidden states, weighted by the attention mask. The pooled representation was passed through a dropout layer with a rate of 0.3 and a linear classification head mapping to the target classes.

Training setup. The model was optimized using AdamW with a learning rate of 2×10^{-5}. Cross-entropy loss was used as the training objective. The model was trained for 3 epochs with a batch size of 16.

5.2 Baseline 2: AraBERT with Class-weighted Loss

Model and tokenizer. We used the pretrained AraBERT model `aubmindlab/bert-base-arabertv02` with its corresponding tokenizer. Tokenization was performed with truncation and padding to a maximum sequence length of 256 tokens.

Model architecture. The model consists of a pretrained AraBERT encoder followed by mean pooling, a dropout layer with a rate of 0.3, and a linear classification head.

Handling class imbalance. To mitigate class imbalance, class weights were computed as:

$$w_i = \frac{N}{C \times n_i}, \tag{1}$$

where N is the total number of training samples, $C = 82$ is the number of classes, and n_i is the number of samples in class i. This weighting scheme corresponds to the `balanced` mode in SCIKIT-LEARN, ensuring that each class contributes equally to the loss regardless of its frequency.

Training setup. All parameters were fine-tuned using AdamW with a learning rate of 1×10^{-5}. The model was trained for up to 8 epochs.

Early stopping. Early stopping was applied based on the validation macro-F1 score with a patience of 2 epochs. The best-performing checkpoint was selected.

5.3 Baseline 3: AraBERT with Arabic Preprocessing and Attention Pooling

Arabic text preprocessing. Prior to tokenization, all texts were normalized using a dedicated Arabic preprocessing pipeline designed for medical and forum-style data. The preprocessing included: (1) removal of question-answer structural markers and forum-specific expressions, (2) normalization of Arabic orthography, (3) removal of Arabic diacritics, and (4) whitespace normalization.

Model architecture. The model is based on a pretrained AraBERT encoder combined with a learnable attention-based pooling mechanism. The attention mechanism computes a weighted sum of token representations:

$$\alpha_i = \mathrm{softmax}(W_a \cdot h_i + b_a), \tag{2}$$

for each token embedding h_i, where W_a and b_a are trainable parameters. The pooled representation is then computed as

$$r = \sum_i \alpha_i \cdot h_i. \tag{3}$$

The pooled representation was passed through a dropout layer with a rate of 0.1 and a linear classification head.

Training setup. Optimization was performed using AdamW with a learning rate of 2×10^{-5} and a cosine learning rate scheduler with 10% warmup steps. Gradient clipping with a maximum norm of 1.0 was applied. The model was trained for up to 6 epochs.

Early stopping and evaluation. Early stopping was applied based on validation macro-F1 score with a patience of 2 epochs.

5.4 Baseline 4: AraBERT with Normalization and Mean Pooling

Arabic text preprocessing. The same Arabic normalization pipeline as in Baseline 3 was applied consistently to all splits.

Model architecture. The model consists of a pretrained AraBERT encoder followed by mean pooling, a dropout layer with a rate of 0.3, and a linear classification head.

Training setup. The model was fine-tuned using AdamW with a learning rate of 1×10^{-5} for up to 8 epochs. Early stopping with a patience of 2 epochs was applied based on validation macro-F1 score.

6 Experimental Results

Table 1 summarizes the performance of all experimental configurations. The strongest validation performance was achieved by Baseline 2, which combines class-weighted loss with longer input sequences and early stopping. While attention pooling introduced additional modeling capacity, it did not consistently outperform simpler mean-pooling baselines.

7 Results

The proposed approach achieved competitive performance, ranking 12th on the private leaderboard with a best macro-F1 score of 0.411. Table 1 summarizes the performance of all experimental configurations evaluated using macro-averaged F1 score.

7.1 Performance Comparison

Among the tested systems, the strongest performance is achieved by Experiment 2: AraBERT

Exp.	Model	Max Len	Class Weights	Norm.	Pooling	Epoch	Batch	LR	Public F1	Private F1
1	AraBERT-twitter	128	No	No	Mean	3	16	2×10^{-5}	0.3670	0.3563
2	AraBERT	256	Yes	No	Mean	7	8	2×10^{-5}	**0.3871**	**0.4114**
3	AraBERT	256	Yes	Yes	Attention	6	8	2×10^{-5}	0.3635	0.3600
4	AraBERT	256	Yes	Yes	Mean	5	8	1×10^{-5}	0.3157	0.3264

Table 1: Performance comparison of AraBERT-based configurations. Max Len = maximum sequence length; Class Weights = class-weighted loss; Norm. = Arabic text normalization; Pooling = representation pooling strategy. All models evaluated using macro-F1 score.

with class-weighted loss and mean pooling over 256-token sequences. This configuration attains a macro-F1 score of 0.387 on the public evaluation set and 0.411 on the private test set. Notably, more complex architectures, such as attention-based pooling (Experiment 3), did not consistently outperform this simpler baseline, achieving 0.360 on the private set despite incorporating Arabic text normalization.

7.2 Analysis of Modeling Choices

The performance differences across experiments reveal several insights. Comparing Experiment 1 (baseline without class weights) and Experiment 2 (with class weights), we observe a substantial improvement of +0.055 private F1. This gain can be attributed to two factors: (1) class-weighted cross-entropy loss, which explicitly rebalances the contribution of rare classes, and (2) increased sequence length from 128 to 256 tokens, allowing the model to capture more context from longer medical consultations.

The underperformance of Experiment 3 (attention pooling + normalization) relative to Experiment 2 (-0.051 private F1) suggests that learnable attention mechanisms may require substantially more training data to converge effectively, or that mean pooling provides more robust aggregation for this task. Similarly, Experiment 4's lower results (-0.085 private F1) despite sharing most design choices with Experiment 2 indicate that the reduced learning rate (1×10^{-5} vs. 2×10^{-5}) combined with early stopping at epoch 5 led to underfitting.

7.3 Error Patterns

While we do not have access to per-class predictions on the test set, the validation set analysis and task characteristics suggest predictable error patterns. Most errors are expected to occur in low-frequency classes (those with fewer than 50 training examples), where limited data prevents the model from learning discriminative representations. Confusion is likely concentrated among semantically related medical specialties.

Experiments were run on Apple Silicon (M4 Max, 48 GB RAM) using the MPS backend.

8 Discussion

Our findings highlight several practical insights for imbalanced Arabic NLP. First, class-weighted loss is more impactful than architectural modifications when training data is severely skewed. Second, longer sequence lengths (256 vs. 128 tokens) provide marginal gains, suggesting that key diagnostic information appears early in medical consultations. Third, attention-based pooling underperformed mean pooling, possibly due to limited training data preventing the attention mechanism from learning robust weights.

If additional development time were available, future work would explore data augmentation strategies, hierarchical label modeling, and domain-adaptive pretraining on larger Arabic medical corpora. From a deployment perspective, achieving reliable performance on rare but clinically important categories remains essential, as misclassification in these cases may have practical consequences. Fairness and bias considerations are also relevant, particularly in ensuring consistent performance across underrepresented medical categories.

9 Conclusion

We presented a systematic study of AraBERT-based models for Arabic medical consultation classification under severe class imbalance. Through extensive experimentation, we showed that class-weighted loss combined with simple pooling strategies outperforms more complex architectural modifications. Our results emphasize the importance of loss design and evaluation metrics for long-tailed Arabic NLP tasks and provide practical insights for building robust medical text classification systems.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11-16 May 2020*.

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), colocated with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.