AbjadMed: Arabic Medical Text Classification at AbjadNLP 2026

Pranav Gupta
Lowe's

Niranjan Kumar M
Lowe's

Balaji Nagarajan
Lowe's

Imed Zitouni
Meta

Mo El-Haj
VinUniversity

Abstract

We present *AbjadMed*, a shared task on Arabic medical text classification organised as part of the 2nd AbjadNLP workshop at EACL 2026. The task targets supervised multi-class classification under realistic conditions of severe class imbalance, fine-grained category structure, and naturally occurring label noise. Participants assign each Arabic medical question–answer instance to one of 82 predefined categories derived from real healthcare consultations. The dataset is based on the Arabic Healthcare Dataset (AHD) and is released as curated training and test splits containing 27,951 and 18,634 instances respectively, while preserving the original label distribution. Systems are evaluated using macro-averaged F1 to emphasise performance on minority medical topics. Results show that Arabic medical text classification remains challenging even with modern pretrained models, particularly for low-frequency and semantically overlapping categories. AbjadMed provides a reproducible benchmark for studying robustness and generalisation in Arabic healthcare NLP.

1 Introduction

Several efforts have sought to address data scarcity in Arabic healthcare NLP by releasing domain-specific datasets. Notably, the Arabic Healthcare Dataset (AHD) provides a large-scale collection of health-related question–answer pairs across a wide range of medical categories, offering a valuable foundation for classification and generation tasks (Al-Majmar et al., 2024). Complementary datasets have focused on narrower settings, such as disease-oriented classification or medical question answering, demonstrating the feasibility of supervised learning in the medical domain but often under controlled or relatively balanced conditions (Hammoud et al., 2021). While these resources have enabled methodological progress, they have also highlighted persistent challenges related to class imbalance and fine-grained category distinctions.

Recent advances in pretrained Arabic language models and domain adaptation techniques have further improved baseline performance on health-related tasks. Work on Arabic medical question answering, including the AraMed dataset and the AraHealthQA shared task, illustrates how classification-related subtasks such as intent detection and topic routing underpin more complex healthcare NLP pipelines (Alasmari et al., 2024; Alhuzali et al., 2025). At the same time, studies on Arabic health text classification report that strong performance on dominant categories does not necessarily translate into robustness across underrepresented medical topics, particularly in realistic clinical or consumer-health settings (Al-Fuqaha'a et al., 2024).

Despite this progress, the lack of shared benchmarks that explicitly prioritise realistic data characteristics has limited systematic comparison across approaches. In particular, few evaluation settings address the combined effects of severe class imbalance, semantically overlapping labels, and naturally occurring annotation noise, all of which are common in real-world healthcare data. As a result, reported improvements often reflect dataset-specific optimisation rather than generalisable advances in Arabic medical NLP.

To address this gap, we introduce *AbjadMed*, a shared task on Arabic medical text classification organised as part of the 2nd AbjadNLP workshop co-located with EACL 2026 (El-Haj, 2025, 2026). The task focuses on multi-class classification of Arabic medical question–answer text into 82 predefined categories. The dataset is intentionally challenging: category frequencies are highly skewed, and the label space includes closely related or overlapping medical topics derived from naturally occurring annotation practices. Systems are evaluated using macro-averaged F1 to ensure that perfor-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 506–514
March 28, 2026. ©2026 Association for Computational Linguistics

mance on minority categories contributes equally to the final score. Beyond leaderboard ranking, AbjadMed aims to provide a common empirical basis for analysing modelling strategies under realistic constraints and for identifying persistent failure modes in Arabic medical text classification.

2 Related Work

Research on Arabic medical and healthcare NLP has expanded in recent years, driven by increased availability of domain-specific datasets and the adoption of pretrained transformer models. Nevertheless, compared to English, Arabic medical text classification remains underexplored, particularly under realistic conditions involving label imbalance and fine-grained category structures. Corpus-based studies of biomedical language further highlight the difficulty of identifying salient and characteristic patterns in heterogeneous medical text, especially when categories are closely related or unevenly distributed (Prentice et al., 2021).

Early work on Arabic text classification established foundational methods using classical machine learning and lexical features (Al-Harbi et al., 2008). While effective for general-domain categorisation, such approaches were limited in their ability to handle domain-specific terminology and long, heterogeneous texts common in medical settings. Subsequent studies introduced specialised datasets for Arabic disease and symptom classification, often framing the task as multi-class or multi-label prediction (Hammoud et al., 2021). These efforts demonstrated the feasibility of supervised medical classification but were typically restricted to smaller label sets of typically less than 20 categories and curated distributions.

More recent work has focused on healthcare-oriented Arabic question answering, which is closely related to classification through intent detection and topic routing. For example, the AraMed dataset introduced a large-scale Arabic medical QA resource built from consumer health questions, enabling systematic evaluation of pretrained Arabic transformer models in the medical domain (Alasmari et al., 2024). Building on this direction, the AraHealthQA 2025 shared task had provided the first standardised evaluation framework for Arabic healthcare question answering, comprising multiple tracks that assess both retrieval-based and reasoning-based capabilities of modern language models (Alhuzali et al., 2025). System description papers from AraHealthQA highlighted the central role of classification-related subtasks, including multiple-choice selection and medical intent recognition, as prerequisites for effective QA pipelines.

Parallel to QA-focused research, several studies have addressed Arabic health text analysis and classification more directly. Al-Fuqaha'a et al. (Al-Fuqaha'a et al., 2024) propose a robust multi-class classification approach for Arabic clinical text, explicitly discussing challenges related to dialectal variation and domain ambiguity. Related work has also examined semantic profiling and entity-centric analysis of biomedical text, demonstrating how structured medical knowledge representations can support large-scale analysis across medical domains (El-Haj et al., 2018; Lal et al., 2025). These findings reinforce the importance of evaluation settings that account for semantic overlap and uneven category distributions.

From a broader perspective, systematic reviews of Arabic text classification research confirm that healthcare remains one of the most challenging application domains due to sparse annotated data, terminology variation, and severe class imbalance (Wahdan et al., 2024). Complementary studies on Arabic health communication further show that linguistic complexity and stylistic variation in patient-facing materials can affect downstream processing and categorisation (Malik et al., 2019). These surveys consistently identify the lack of large, openly evaluated benchmarks as a limiting factor for progress.

The AbjadMed shared task we introduce in this paper is positioned within this landscape as a supervised Arabic medical text classification benchmark that emphasises realistic data characteristics, given the importance of classification and open-source benchmarks. Unlike prior work that focuses primarily on question answering or coarse-grained disease categorisation, AbjadMed targets fine-grained category prediction across a large label space, using macro-averaged evaluation to prioritise performance on underrepresented medical topics. In doing so, it complements existing Arabic healthcare QA initiatives by isolating and rigorously evaluating a core classification capability that underpins triage, routing, and decision-support systems.

3 Task Description

3.1 Task Definition

The AbjadMed task is formulated as a single-label, multi-class classification problem. Given an input text instance drawn from the medical domain, the system must assign exactly one label from a fixed set of 82 categories. Each instance consists of a question–answer pair written in Arabic and provided as a single textual input.

Let $x_i \in \mathcal{X}$ denote an input text and $y_i \in \{0, \ldots, 81\}$ its corresponding gold label. The objective is to learn a function $f : \mathcal{X} \to \{0, \ldots, 81\}$ such that $\hat{y}_i = f(x_i)$ approximates y_i as accurately as possible under the evaluation metric defined below.

3.2 Dataset Structure

The released dataset follows a tabular format with three fields:

- `text`, containing the Arabic medical question–answer text;
- `category`, providing the English name of the medical category;
- `label`, an integer identifier corresponding to the target class.

Category names were originally defined in Arabic and subsequently translated into English using a large language model to support inspection and analysis. The prediction target is the integer `label` field; the `category` field is provided as auxiliary information only.

Two characteristics of the dataset are particularly relevant. First, the distribution of instances across categories is highly skewed, with a small number of frequent classes and a long tail of sparsely represented categories. Second, the label set reflects natural annotation practices and therefore includes semantically adjacent or partially overlapping categories. Evaluation is performed strictly with respect to the original labels, without consolidation or post-hoc smoothing.

3.3 Evaluation Metric

System performance is assessed using macro-averaged F1 over all 82 categories. For each class c, an F1 score $\mathrm{F1}_c$ is computed in a one-vs-all setting, and the final score is obtained by averaging across classes:

$$\mathrm{MacroF1} = \frac{1}{82} \sum_{c=0}^{81} \mathrm{F1}_c. \qquad (1)$$

This metric assigns equal weight to each category, regardless of frequency, and therefore penalises systems that perform poorly on minority classes even if overall accuracy is high.

3.4 Submission Protocol

Participants submit predictions in CSV format with two columns: `Id`, identifying the input instance, and `Predicted`, containing the predicted integer label. Submissions must adhere strictly to the provided format, including preservation of row order and exclusion of any index column.

4 Dataset

The data used in the AbjadMed shared task are derived from the Arabic Healthcare Dataset (AHD) introduced by Al-Majmar et al. (Al-Majmar et al., 2024). AHD is a large-scale Arabic medical question–answer corpus collected from the Altibbi medical platform and released in raw form without linguistic pre-processing. The full dataset contains more than 808,000 question–answer pairs spanning 90 medical categories and represents, to date, one of the most comprehensive publicly available Arabic healthcare resources.

For the purposes of the shared task, a curated subset of AHD was constructed to enable controlled evaluation under realistic but tractable conditions. The subset focuses on supervised medical text classification and retains the natural properties of the source data, including class imbalance, heterogeneous text length, and semantically overlapping category labels. No manual relabelling, category merging, or linguistic normalisation was applied beyond the selection of instances and categories described below.

4.1 Data Selection and Scope

The shared-task dataset comprises 46,585 Arabic medical question–answer instances, split into 27,951 training examples and 18,634 test examples. The subset covers 82 medical categories selected from the original AHD label space. To avoid extreme data imbalance, we downsampled highly frequent categories to a maximum of 600 examples per category. 8 rare categories were excluded to ensure minimal learnability while preserving the long-tailed distribution characteristic of real-world medical data. Table 5 in Appendix A shows the training set complete categories distribution.

Each instance corresponds to a medical consultation consisting of a user question and a professional response, concatenated and provided as a single text field. Category labels were inherited directly from the source platform. While category names were originally provided in Arabic, they were translated into English using a large language model to support modelling and interpretation; participants were required to predict the corresponding integer label.

Table 1 summarises the main statistics of the dataset used in the shared task.

Statistic	Train	Test
Number of question-answer pairs	27,951	18,634
Average words per question-answer pair	59.00	58.28
Minimum words per question-answer pair	8	9
Maximum words per question-answer pair	2,223	1,886
Number of labels	82	–

Table 1: Summary statistics for the training and test datasets used in the AbjadMed shared task.

4.2 Text Characteristics

The dataset exhibits substantial variation in text length. While the average instance contains approximately 59 words, the maximum length exceeds 2,000 words in the training set, reflecting detailed medical explanations and follow-up advice. This wide length distribution poses challenges for standard transformer-based architectures and motivates exploration of truncation strategies, long-context modelling, and hierarchical representations.

The text is primarily written in Modern Standard Arabic, with occasional dialectal expressions, numerals, medical abbreviations, and non-Arabic symbols, consistent with the properties reported for the full AHD corpus (Al-Majmar et al., 2024). Diacritics are rare, and no spelling normalisation or token-level cleaning was performed. An example from the train split of the dataset is given below, with Arabic text transliterated for readability:

text

al sual

—

as salam alaykum ana musab bi faqr al dam al manjali al siklsil ilman bi anna nisbat al siklsil 72 fa indama tusbih nisbat al dam 7 fa inna al alam tati bi kathrah fa ma al hall li ziyadat nisbat al dam wa ma al hall li ilaj

al jawab

—

al hall bi al ibtiad an al radrad al nafsiyyah wa taqwiyat al manaah wa tanawul himyah ghidhaiyyah mutawazinah ghaniyyah bi al hadid wa inda huduth nawabat alam sababha wa nuqs hadd bi al khidab al damawi la yujad illa tawid al dam al naqis bi naql al dam

category

Hematological diseases

label

33

English translation generated by GPT-5 (mentioned here for reference purposes only, not a part of the actual dataset)

Question

Peace be upon you. I have sickle cell anaemia, knowing that my sickle cell percentage is 72%. When my hemoglobin level drops to 7, the pain becomes very frequent. What is the solution to increase my blood level, and what is the treatment?

Answer

The solution is to avoid psychological stress, strengthen the immune system, and follow a balanced diet rich in iron. When pain crises occur due to a severe deficiency in hemoglobin, there is no option other than compensating for the missing blood by blood transfusion.

4.3 Label Distribution and Noise

As in the original AHD dataset, category frequencies in the AbjadMed subset are highly imbalanced. A small number of high-level medical topics account for a large proportion of instances, while many specialised categories are represented by relatively few examples. In addition, the category system reflects organically evolved labels from the source platform, resulting in semantically overlapping or closely related categories (e.g., multiple dental or reproductive health topics).

Evaluation strictly follows the original labels, and no soft matching or label equivalence is assumed. This design choice intentionally exposes systems to realistic annotation noise and boundary ambiguity, encouraging robustness rather than optimisation for artificially clean label sets.

4.4 Train–Test Split and Evaluation Use

The training split includes both text and labels and was released to participants for model development.

The test split contains only text and instance identifiers, with labels withheld and used exclusively for evaluation on the competition platform. Leaderboard scores are computed using macro-averaged F1 over all 82 categories, assigning equal weight to each class and penalising models that ignore low-frequency labels.

By retaining the natural imbalance, label overlap, and length variability of Arabic medical consultations, the AbjadMed dataset provides a challenging and ecologically valid benchmark for Arabic medical text classification.

5 Results

This section summarises the performance of submitted systems on the AbjadMed shared task, based on the official Kaggle leaderboards. Evaluation was conducted using macro-averaged F1 over 82 classes, computed on a hidden test set that was split internally by the platform into public and private partitions.

5.1 Participation Overview

The shared task attracted strong engagement from the community. In total, 61 individuals registered for the competition, forming 40 teams and submitting 334 runs over the evaluation period. This level of participation reflects sustained interest in Arabic medical NLP and highlights the relevance of classification tasks that combine domain specificity with realistic data challenges such as class imbalance and label noise.

5.2 Leaderboard Results

Tables 2 and 3 report the top-performing systems on the private and public leaderboards, respectively. Final rankings are determined solely by the private leaderboard, which is computed over approximately half of the hidden test set and remains inaccessible to participants during the competition. Given that the public and private leaderboard datasets each have roughly 50% of the 18,634 non-training examples, we expect an uncertainty of $\sim \frac{1}{\sqrt{9317}} = O(0.01)$ in the reported model performances.

5.3 Performance Trends

Overall performance levels indicate that Arabic medical text classification at fine-grained category resolution remains a challenging problem. Even the best-performing system achieves a macro-F1 below 0.70 on the private leaderboard, with a noticeable performance drop compared to public leaderboard scores. This gap suggests some degree of overfitting to the public split or sensitivity to topic distribution shifts between the two partitions.

Rank	Team	Macro-F1
1	F.A.H	0.6732
2	Gleb Shanshin	0.5139
3	HCMUS_PrompterXPrompter	0.4902
4	HCMUS_FanALong	0.4862
5	boy Magic	0.4611
6	REGLAT	0.4606
7	baellouf	0.4398
8	Yuchen Liu	0.4341
9	MedArabs	0.4219
10	DerivedByData	0.4192

Table 2: Top systems on the private leaderboard, ranked by macro-averaged F1.

Rank	Team	Macro-F1
1	F.A.H	0.7422
2	Gleb Shanshin	0.5071
3	HCMUS_FanALong	0.4619
4	REGLAT	0.4615
5	HCMUS_PrompterXPrompter	0.4570
6	boy Magic	0.4475
7	Yuchen Liu	0.4245
8	baellouf	0.4144
9	KvochurHegel	0.4087
10	MedArabs	0.4068

Table 3: Top systems on the public leaderboard, based on a visible subset of the test data.

The spread of macro-F1 scores across teams is relatively narrow beyond the top-ranked system, with many submissions clustering between 0.35 and 0.50. This pattern reflects the difficulty of achieving consistent gains across all 82 categories, particularly for minority and overlapping classes. Table 2 and Table 3 describe the scores for the top-performing teams on the private and public leaderboards respectively.

5.4 Discussion and Limitations

Several factors likely contribute to the observed performance ceiling. These include severe class imbalance, semantic overlap between categories, and substantial variation in input length. Together, these factors make consistent performance across all classes difficult and highlight the need for further research on robust modelling strategies for Arabic medical text.

The base models finetuned by participants included CamelBERT (Inoue et al., 2021), MARBERT (Abdul-Mageed et al., 2021), ARBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020) and Qwen3-8B (Yang et al., 2025). We also observed a range of techniques in the submissions, including class weights, QLoRA, attention pooling, kNN-based retrieval, LDAM loss, adversarial training, back-translation, focal loss, mean pooling, and custom multi-layer perceptron classification heads. We further describe the methods used by some of the participating teams below, and summarize them in Table 4.

F.A.H. utlilized data augmentation and trained an XGBoost model with sample weighting, by using embeddings generated from the AraBERT model as its feature vector.

ArabicMedicalBERT-QA-82 (Shanshin, 2026) fine-tuned an AraBERT-based medical classifier with strong class reweighting and extensive data augmentation, using 10-fold ensembling to stabilize performance under extreme label imbalance. Their approach leveraged a domain-specific pretrained backbone that already encoded the 82-class structure.

baellouf (Khallouf, 2026) employed Qwen3-8B fine-tuned with all-linear QLoRA, combining large-scale instruction tuning with a Dice+CE hybrid loss and heavy data augmentation via machine-translated medical QA data, substantially expanding the training corpus. This brought LLM-scale capacity to the Arabic medical classification setting.

GATech (Khamis, 2026) evaluated a wide range of encoder models and selected AraBERT as the most robust backbone, enhancing it with mean-plus-attention pooling and multisample dropout. Although LLM-based reranking was explored, a pure encoder-based AraBERT system achieved the strongest results.

HCMUS_PrompterXPrompter (Dao Sy et al., 2026) proposed a hybrid classification–retrieval framework aimed at taming the long tail in Arabic medical text, combining prompting and retrieval strategies to improve coverage of rare classes. The system focused on integrating semantic search with supervised classification.

KvochurHegel (Le, 2026) combined LDAM loss with adversarial training to explicitly address class imbalance in Arabic medical QA classification. The approach targeted margin-aware optimization to improve minority-class separation under extreme skew.

MedArabs (Singh, 2026) explored data- and algorithm-level fusion, combining multiple models and augmentation strategies to improve robustness under imbalance. Their system emphasized ensemble-style integration across representations and training regimes.

MetaSwarm (Jaisy, 2026) introduced a class-balanced discovery and optimization pipeline tailored to medical diglossia in Abjad scripts. The method focused on forensic data handling and imbalance-aware optimization for fine-grained Arabic medical categories.

Olga Snissarenko (Snissarenko, 2026) fine-tuned AraBERT with mean pooling instead of the CLS token, using dynamically balanced class weights and early stopping. The system achieved strong macro-F1 through careful regularization and imbalance-aware training.

REGLAT (Fetouh et al., 2026) adopted a hierarchical architecture in which a fine-tuned Arabic BERT produces embeddings consumed by a KNN classifier, with a specialist MLP correcting rare-class predictions. This hybrid BERT–KNN–MLP design explicitly targets minority labels through selective augmentation and hierarchical correction.

REIGNITE (Rifat and Dewan, 2026) combined aggressive minority-class augmentation with imbalance-aware fine-tuning and model ensembling. Predictions from CAMeLBERT and multiple AraBERT variants were merged via majority voting under a class-weighted focal loss.

Supachoke (Nguyen et al., 2026) fine-tuned AraBERT with Arabic-specific normalization and weighted cross-entropy, using mixed-precision training and early stopping for stability. The system emphasized clean preprocessing and efficient transformer optimization.

Sujith Kanakkassery (Kanakkassery, 2026) fine-tuned MARBERT with a custom MLP classification head on top of the CLS representation, using class-weighted loss and label smoothing. Careful training control and preprocessing improved minority-class performance under severe imbalance.

Tashkees-AI (Eldin, 2026) implemented a flat MARBERTv2-based classifier, finding it superior to hierarchical and RAG-based alternatives due to error propagation in multi-stage setups. Strong preprocessing and weighted loss were central to handling the 82-way imbalance.

Team Name	Description of the team's best performing model
F.A.H.	XGBoost model using AraBERT as a frozen feature extractor
ArabicMedicalBERT-QA-82	AraBERT fine-tuning with class weights (10-fold)
baellouf	Qwen3-8B with QLoRA (all-linear)
GATech	AraBERT encoder with attention pooling
HCMUS_PrompterXPrompter	Hybrid AraBERT + kNN retrieval system
KvochurHegel	ARBERTv2 with LDAM loss and adversarial training
MedArabs	AraBERT with back-translation and class-balanced loss
MetaSwarm	CAMeLBERT with class-balanced focal loss
Olga Snissarenko	AraBERT with mean pooling
REGLAT	Hierarchical BERT + KNN + MLP
REIGNITE	Ensemble of Arabic BERTs with focal loss
Supachoke	AraBERT fine-tuning with weighted loss
Sujith Kanakkassery	MARBERT with custom MLP head
Tashkees-AI	MARBERTv2 flat classifier

Table 4: Summary of the methods used by a selection of teams in this shared 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. Association for Computational Linguistics.

Shrouq Al-Fuqaha'a, Nailah Al-Madi, and Bassam Hammo. 2024. A robust classification approach to enhance clinic identification from arabic health text. *Neural Computing and Applications*, 36(13):7161–7185.

S Al-Harbi, A Almuhareb, A Al-Thubaity, M. S. Khorsheed, and A Al-Rajeh. 2008. Automatic arabic text classification. In *Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data (01/03/08)*.

Nashwan Ahmed Al-Majmar, Hezam Gawbah, and Akram Alsubari. 2024. Ahd: Arabic healthcare dataset. *Data in Brief*, 56:110855.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. Aramed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 50–56.

Hassan Alhuzali, Farah E Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, and 1 others. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pages 9–15. European Language Resource Association.

Duy Minh Dao Sy, Trung Kiet Huynh, Nguyen Dinh Ha Duong, Nguyen Chi Tran, Phu Quy Nguyen Lam, and Hoa Phu Pham. 2026. When Classification Meets Retrieval: Taming the Long Tail in Arabic Medical Text Classification. 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.

Mahmoud El-Haj, Paul Rayson, Scott SL Piao, and Jo Knight. 2018. Profiling medical journal articles using a gene ontology semantic tagger. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mo El-Haj. 2025. Proceedings of the 1st workshop on nlp for languages using arabic script (abjadnlp 2025). In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script, held at COLING 2025*, Abu Dhabi, United Arab Emirates.

Mo El-Haj. 2026. Proceedings of the 2nd workshop on nlp for languages using arabic script (abjadnlp 2026). In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script, held at EACL 2026*, Rabat, Morocco.

Fatimah Mohamed Emad Eldin. 2026. Flat vs. Hierarchical Classification for Fine-Grained Arabic Medical QA. 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 Megahed Fetouh, Mohammed Rahmath, Omer Dawood, Mariam Labib, Nsrin Ashraf, and Hamada Nayel. 2026. Handling Imbalanced Arabic Medical Text Classification via Hierarchical KNN-MLP Architecture. 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.

Jaafar Hammoud, Aleksandra Vatian, Natalia Dobrenko, Nikolai Vedernikov, Anatoly Shalyto, and Natalia Gusarova. 2021. New arabic medical dataset for diseases classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 196–203. Springer.

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 92–104. Association for Computational Linguistics.

Rahul Jaisy. 2026. Forensic Optimization and Class-Balanced Discovery for Medical Diglossia in Abjad

Scripts. 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.

Sujith Kanakkassery. 2026. Imbalance-Aware Transformer Fine-Tuning for Arabic Medical Text Classification. 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.

Abdallah Khallouf. 2026. Efficient Fine-Tuning with All-Linear LoRA for Arabic Medical QA Classification. 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 Khamis. 2026. Bidirectional Encoders vs. Causal Decoders: Insights from 82-Class Arabic Medical Classification. 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.

Daisy Monika Lal, Paul Rayson, Christopher Peter, Ignatius Ezeani, Mo El-Haj, Yafei Zhu, and Yufeng Liu. 2025. Lens: Learning entities from narratives of skin cancer. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 20–27.

Minh-Hoang Le. 2026. Combining LDAM Loss and Adversarial Training for Arabic Medical Question-Answer Classification. 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.

Abdulaziz Malik, Mahmoud El-Haj, and Michael K Paasche-Orlow. 2019. Readability of patient educational materials in english versus arabic. *HLRP: Health Literacy Research and Practice*, 3(3):e170–e173.

Thanh Phu Nguyen, Tuan Thai Huy Nguyen Cu, Son Thai Pham, and Tri Duy Ho Nguyen. 2026. Enhancing Arabic Medical Text Classification Using Fine-Tuned AraBERT. 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.

Sheryl Prentice, Jo Knight, Paul Rayson, Mahmoud El Haj, and Nathan Rutherford. 2021. Problematising characteristicness: a biomedical association case study. *International Journal of Corpus Linguistics*, 26(3):305–335.

Nahid Montasir Rifat and Foyez Ahmed Dewan. 2026. Imbalance-Aware Fine-Tuning of Pretrained Arabic Transformers for Arabic Medical Text Classification Task. 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.

Gleb Shanshin. 2026. Fighting Class Imbalance in Arabic Medical Text Classification. 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.

Amrita Singh. 2026. Arabic Medical Text Classification via Data- and Algorithm-Level Fusion. 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.

Olga Snissarenko. 2026. Arabic Clinical Text Classification with AraBERT: Results from the AbjadMed Shared Task. 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.

Ahlam Wahdan, Mostafa Al-Emran, and Khaled Shaalan. 2024. A systematic review of arabic text classification: areas, applications, and future directions. *Soft Computing-A Fusion of Foundations, Methodologies & Applications*, 28(2).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 8 others. 2025. Qwen3 technical report. *arXiv*, abs/2505.09388.

A Appendix

ID	Category	#	ID	Category	#
0	Addiction	600	41	Internal medicine diseases	600
1	Allergy	333	42	Jaw and dental surgery	411
2	Alternative medicine	232	43	Laboratory	134
3	Anatomy	37	44	Medical services	46
4	Anesthesiology	34	45	Medicinal herbs	241
5	Benign and malignant tumors	600	46	Men's health	600
6	Biochemistry	7	47	Mental health	600
7	Biology	29	48	Microbiology	29
8	Cardiothoracic surgery	345	49	Musculoskeletal and joint diseases	600
9	Cardiovascular diseases	600	50	Neurological diseases	600
10	Chemistry	11	51	Neurosurgery	356
11	Child health	600	52	Nutrition	600
12	Congenital malformations	13	53	Optometry	49
13	Dental diseases	600	54	Oral diseases	600
14	Dental health	600	55	Orthopedic surgery	600
15	Dentistry	600	56	Pathology	34
16	Dermatological diseases	600	57	Pediatric diseases	600
17	Diabetes	600	58	Pediatric surgery	10
18	Diagnosis	154	59	Pharmacology	600
19	Ear, nose, and throat (ENT)	600	60	Physiology	20
20	Embryology	40	61	Physiotherapy	251
21	Endocrine diseases	600	62	Plastic surgery	600
22	Eye diseases	600	63	Pregnancy and childbirth	600
23	First aid	41	64	Preventive medicine	20
24	Gastrointestinal diseases	600	65	Psychiatric diseases	600
25	General medicine	600	66	Psychology	229
26	General surgery	600	67	Public health	600
27	Genetic diseases	100	68	Radiology	50
28	Genetics	26	69	Ramadan	8
29	Geriatric health	10	70	Respiratory diseases	600
30	Gynecologic surgery	156	71	Rheumatic diseases	20
31	Gynecological diseases	600	72	Sexual health	600
32	Health and sports	600	73	Sexually transmitted diseases	600
33	Hematological diseases	600	74	Skin and beauty	600
34	History of medicine	11	75	Toxicology	33
35	Hormones	144	76	Urogenital diseases	600
36	Hypertension	600	77	Urological surgery	235
37	Immunology	55	78	Vaccines and immunizations	19
38	In vitro fertilization (IVF)	7	79	Vascular surgery	7
39	Infectious diseases	242	80	Vitamins and minerals	90
40	Infertility	232	81	Women's health	600

Table 5: Label–category mapping and number of training instances per category