Supachoke at AbjadMed: Enhancing Arabic Medical Text Classification Using Fine-Tuned AraBERT

Nguyen Phu Thanh, Cu Nguyen Huy Thai Tuan, Pham Thai Son, Nguyen Ho Duy Tri
University of Information Technology (UIT), VNU-HCM
Ho Chi Minh City, Vietnam
23521452@gm.uit.edu.vn, 23521706@gm.uit.edu.vn
23521361@gm.uit.edu.vn, trinhd@uit.edu.vn

Abstract

Medical text classification is an important task in healthcare NLP, yet Arabic medical texts remain underexplored due to linguistic complexity and limited annotated data. In this paper, we study the effectiveness of AraBERT, a pre-trained Arabic transformer model, for Arabic medical text classification. We fine-tune AraBERT on a labeled medical dataset and evaluate its performance using standard classification metrics. Experimental results show that our fine-tuned AraBERT model achieves a private leaderboard score of 0.4076 and ranks 13th among participating teams, outperforming classical machine learning baselines and other transformer variants. These findings highlight the potential of transformer-based approaches for Arabic medical NLP and motivate further research.

1 Introduction

Medical text classification is a promising task in Natural Language Processing (NLP) (Lan, 2026), enabling applications such as clinical decision support, medical document indexing, and health information retrieval. In this work, we focus on Arabic medical text classification, a relatively underexplored area due to the linguistic complexity of Arabic and the limited availability of annotated medical data. We address this task as defined in the shared task overview paper (Gupta et al., 2026), which formulates medical text classification as a supervised learning problem over Arabic texts.

Our system is based on AraBERT (Antoun et al., 2021), a pre-trained transformer model specifically designed for Arabic. We adopt a fine-tuning strategy in which AraBERT is adapted to the medical domain using the provided labeled training data. This approach allows the model to capture both contextual semantics and domain-specific medical terminology while handling the rich morphology of Arabic.

We evaluate our system using the official task metrics and compare it with baseline methods. The AraBERT-based model attains competitive performance, exceeding traditional machine learning and non-contextual embedding approaches. Our best submission ranked 13th on the official test set.

Our contributions include: (1) applying fine-tuned AraBERT to Arabic medical QA texts, (2) benchmarking against other Arabic transformers (SaudiBERT, CamelBERT) to validate model selection, and (3) mitigating class imbalance with weighted loss. To support reproducibility, we release our implementation at: https://github.com/PhuThanh3012/Supachoke_AbjadNLP.

2 Related Work

Machine Learning Approaches Various classical machine learning paradigms have been explored to address the linguistic complexities of Arabic Text Classification. Hamood (Hamood et al., 2014) introduced an improved k-Nearest Neighbor (k-NN) algorithm, showing that modifications to distance calculation can yield higher accuracy. Additionally, Harrag (Harrag and El-Qawasmah, 2009) explored the efficacy of Artificial Neural Networks (ANN), highlighting that neural architectures can effectively capture non-linear relationships within Arabic text, providing a robust alternative to traditional statistical learners.

Pre-trained Language Models Prior work on Arabic NLP has demonstrated the effectiveness of pretrained models such as Arabic-BERT (Safaya et al., 2020), MARBERT (Abdul-Mageed et al., 2021), SaudiBert (Qarah, 2024) and CamelBERT (Inoue et al., 2021). A comprehensive review by Alammary (Alammary, 2022) synthesized 48 studies, reporting that Arabic-specific BERT variants such as AraBERT and MARBERT consistently outperform Multilingual BERT (Devlin et al., 2018) due to richer pretraining corpora and broader cov-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 127–131
March 28, 2026.

erage of dialectal Arabic. Most prior works adopt task-specific fine-tuning and achieve strong results comparable to English BERT on analogous tasks, particularly for social media and news text.

3 Background

This work is conducted within the framework of the **EACL 2026 Abjad NLP Shared Task: Medical Text Classification in Arabic**. The task is formulated as a supervised multi-class text classification problem where each input instance is mapped to exactly one label among **82** predefined classes.

3.1 Task Description

Each sample consists of a short medical question-answer pair written in Arabic. These pairs typically describe patient symptoms, disease histories, treatment concerns, or health-related advice. The input is the text segment, and the output is an integer label. For example, the input: اذا مكيلع مالسلا يلجنملا مدلا رقفب باصم ... is associated with *Hematological diseases* (label 33), while: پاش انا مياذ اذو ىدارا لا لوبق ىدنعو هنس ٢٥ ىدنع ... is mapped to *Urogenital diseases* (label 76).

3.2 Challenges

Arabic medical QA classification is significantly more challenging than generic text classification (Alrayzah et al., 2023). First, each instance contains a paired question and answer, requiring models to capture cross-segment relationships. Second, the domain is highly specialized with medical terminology, while the input text often contains dialectal variations and spelling inconsistencies common in user-generated content. Furthermore, Arabic's rich morphology (root-and-pattern system) increases data sparsity. Finally, the label space is fine-grained with semantic overlap (e.g., distinguishing between different types of infections), and the data suffers from pronounced class imbalance, where some diseases appear frequently while others are rare.

4 Methodology

We employ two classification approaches: (i) a baseline using classical machine learning, and (ii) a transformer-based model via fine-tuning.

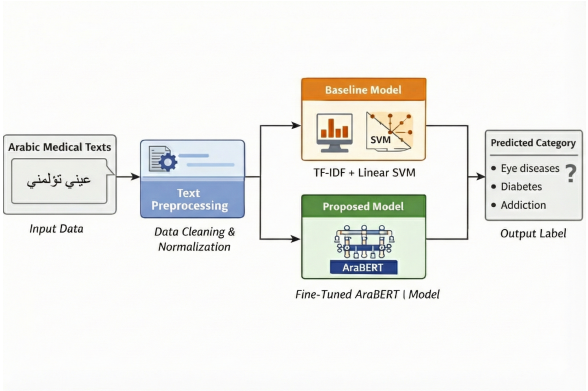

Figure 1: Overview of the proposed pipeline.

4.1 Baseline Approach: TF-IDF + Linear SVM

In the baseline architecture, each input document is transformed into a sparse numerical vector using TF-IDF representation. The resulting vectors are provided as input to a linear Support Vector Machine (SVM) classifier (Dadgar et al., 2016). This baseline establishes whether simple lexical features suffice for this fine-grained medical classification task or whether contextualized representations are necessary.

4.2 Fine-tuning AraBERT Strategy

4.2.1 Architecture Overview

We employ the `bert-base-arabertv02` architecture (Antoun et al., 2021). This model consists of 12 transformer layers, 768 hidden dimensions, and 12 attention heads, totaling approximately 135 million parameters. Unlike multilingual models, AraBERT is pre-trained exclusively on large-scale Arabic corpora. This enables it to better capture Arabic-specific linguistic phenomena, such as the complex agglutinative morphology and orthographic variations, which are crucial for understanding medical texts containing a mix of formal terminology and colloquial expressions.

4.2.2 Model Adaptation and Fine-tuning

The input text is tokenized using AraBERT's WordPiece tokenizer. The contextualized representation of the special `[CLS]` token from the final layer is passed to a linear classification head to produce logits over the 82 categories.

We fine-tune both the transformer encoder and the classification head jointly in an end-to-end manner. This allows the model to adapt its internal representations to capture medical domain knowl-

edge while maintaining linguistic understanding. Fine-tuning handles morphological variations and spelling inconsistencies effectively without additional task-specific architectural components.

5 Experimental Setup

5.1 Preprocessing

We first manually remove dataset-specific markers (e.g., “”, “”). Then, we use `ArabertPreprocessor` to remove HTML markup, replace URLs/emails with placeholders, strip diacritics (tashkeel) and elongation (tatweel), and normalize characters. This reduces orthographic variability while preserving medically meaningful tokens.

5.2 Data Splitting

The provided training set is partitioned into training and validation subsets. We perform a stratified split to preserve the original class distribution across splits, allocating 90% of the data for training and 10% for validation. Stratification is essential due to the pronounced class imbalance (Stone, 2014) across the 82 categories, ensuring that minority classes are represented in both subsets. The validation set is used exclusively for model selection and early stopping, while the official test set is reserved for final leaderboard submission.

5.3 Training Strategy

5.3.1 Class Weights

To address class imbalance, we use balanced class weights computed as:

$$w_k = \frac{N}{K \cdot n_k}$$

where N is the total samples, K is the number of classes, and n_k is the sample count for class k. These weights are incorporated into the weighted cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{K} w_k \, \mathbf{1}[y = k] \, \log\left(\text{softmax}(\mathbf{z})_k\right)$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\mathbf{z}$ denotes the logits.

5.3.2 Hyperparameter Configuration

We fine-tune using AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $5 \cdot 10^{-5}$. Training is carried out for a maximum of 8 epochs with **early stopping** (patience set to 3). To optimize resource usage, we employ **gradient accumulation** (steps=2) and **mixed-precision (FP16)** training, allowing for a larger effective batch size while fitting within GPU memory constraints.

6 Results

6.1 Model Performance and Comparisons

Table 1 presents the experimental results. We compare our proposed AraBERT system against the TF-IDF baseline, two other Arabic-specific transformers (SaudiBERT, CamelBERT), and an unweighted AraBERT variant.

Model	Private Score
TF–IDF + SVM (Baseline)	0.3508
SaudiBERT ($lr = 2e^{-5}$)	0.3570
CamelBERT ($lr = 2e^{-5}$)	0.3823
AraBERT (w/o Class Weights)	0.3759
AraBERT (Proposed System)	**0.4076**

Table 1: Performance comparison on the official test set.

The results yield three key observations:

1. Contextualization Matters: All transformer models outperform the TF-IDF baseline, confirming the necessity of deep contextual representations for this task.

2. Model Selection: Among the pre-trained models, AraBERT achieves superior performance compared to SaudiBERT (0.3570) and CamelBERT (0.3823). We hypothesize that SaudiBERT, which is heavily trained on social media content (Twitter), may struggle with the mixed formal-informal nature of medical QA pairs compared to AraBERT's broader training corpus (news, wikipedia). The medical domain requires a grasp of formal terminology which is less prevalent in short social media posts.

3. Impact of Class Weights: The ablation study shows that removing class weights drops AraBERT's performance to 0.3759. Notably, while the unweighted AraBERT slightly lags behind CamelBERT, the introduction of class weights boosts it to the top performance (0.4076), validating the effectiveness of our loss weighting strategy in handling imbalanced medical data.

6.2 Discussion

Our team, **Supachoke**, ranked **13th** overall with a score of **0.4076** (Table 2).

Team	Rank	Score
F.A.H	1	0.6732
Gleb Shanshin	2	0.5139
HCMUS_PrompterXPrompter	3	0.4902
Our team (Supachoke)	**13**	**0.4076**

Table 2: Leaderboard performance comparison.

While our system demonstrates a solid improvement over baselines, a substantial performance gap of **0.2656** remains between our best run and the top-performing system (0.6732). We attribute this to three primary factors:

1. Lack of Domain Adaptation: Our model was fine-tuned directly on the provided small dataset. In contrast, top systems likely employed **Domain-Adaptive Pretraining (DAPT)**, where the language model is first continued-pretrained on a large corpus of unlabeled Arabic medical texts (e.g., medical articles, health forums) before fine-tuning. This allows the model to learn the nuances of medical terminology which are sparse in general-domain corpora.

2. Single-Model Limitation: We relied on a single AraBERT checkpoint. Leaderboard winners typically utilize **ensemble strategies**, combining predictions from diverse architectures (e.g., ensembling AraBERT with MARBERT and XLM-R) to reduce variance and improve generalization on rare classes.

3. Lexical Ambiguity in QA Pairs: Error analysis suggests our model struggles with semantically similar categories (e.g., distinguishing between different *Infectious diseases*). By treating the Question and Answer as a single flat sequence, the model may rely excessively on keyword matching rather than understanding the diagnostic relationship between the reported symptoms in the question and the medical advice in the answer.

7 Conclusion

In this paper, we presented an effective baseline for Arabic medical text classification using a fine-tuned AraBERT model with a class-weighted loss function. Our approach achieved a private leaderboard score of 0.4076, significantly outperforming the TF-IDF baseline (0.3508) and other Arabic transformer variants like SaudiBERT (0.3570).

Despite these promising results, the performance gap with state-of-the-art systems highlights the limitations of standard fine-tuning on small, imbalanced datasets. Future work will focus on: (1) curating a large-scale unlabeled Arabic medical corpus for Domain-Adaptive Pretraining, (2) implementing a voting ensemble of multiple transformer backbones, and (3) exploring hierarchical classification architectures to better model the dependencies between fine-grained medical categories.

Acknowledgements

We would like to thank Lecturer Nguyen Ho Duy Tri for reviewing this paper. His constructive comments and guidance helped us correct several issues and improve the clarity and quality of the manuscript.

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

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.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: A systematic review. *Applied Sciences*, 12(11).

Asmaa Alrayzah, Fawaz Alsolami, and Mostafa Saleh. 2023. Challenges and opportunities for arabic question-answering systems: current techniques and future directions. *PeerJ Computer Science*, 9:e1633.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. 2016. A novel text mining approach based on tf-idf and support vector machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

KHALED Wail Hamood, AL-SARRAYRIH Haytham Saleem, and KNIPPING Lars. 2014. Arabic text categorization using improved k-nearest neighbour algorithm. *Journal of Applied Computer Science & Mathematics*, 8(3):9–12.

Fouzi Harrag and Eyas El-Qawasmah. 2009. Neural network for arabic text classification. In *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, pages 778–783. IEEE.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Man Lan. 2026. Natural language processing and text mining. In *Artificial Intelligence for Drug Design*, pages 189–215. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Faisal Qarah. 2024. Saudibert: A large language model pretrained on saudi dialect corpora. *arXiv preprint arXiv:2405.06239*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Eric A Stone. 2014. Predictor performance with stratified data and imbalanced classes. *Nature methods*, 11(8):782–783.