REGLAT at AbjadMed: Handling Imbalanced Arabic Medical Text Classification via Hierarchical KNN-MLP Architecture

Ahmed M. Fetouh[1] **Mohammed Rahmath**[2] **Omer Dawood**[2] **Mariam Labib**[3]
Nsrin Ashraf[3] and **Hamada Nayel**[1,2]
[1]Department of Computer Science, Faculty of Computers and AI, Benha University, Egypt
[2]Department of Computer Engineering and Information, College of Engineering,
Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia
[3]Computer Engineering, Elsewedy University of Technology, Cairo, Egypt
ahmed.megahed@fci.bu.edu.eg

Abstract

In this paper, we demonstrate the system submitted to the shared task of medical text classification in Arabic. We proposed a single-model approach based on fine-tuned LLM-based embedding combined with hierarchical classical classifiers, achieving a competitive macro F1-score of 0.46 on the blind test set. We explored various modeling strategies, including tree-based ensembles, LLM, and hierarchical correction for rare classes, highlighting the effectiveness of domain-specific fine-tuning in low-resource settings. The results demonstrate that a single fine-tuned Arabic BERT variant can serve as a strong baseline in extreme imbalance scenarios, outperforming more complex ensembles in simplicity and reproducibility.

1 Introduction

Medical text classification is the automated process of assigning predefined categories to medical-related textual data—such as clinical notes, patient questions, medical abstracts, or diagnostic statements—based on their semantic meaning. It is a core task in medical natural language processing (NLP), enabling machines to interpret specialized medical terminology, irregular grammar, and domain-specific expressions found in healthcare texts (Wang et al., 2023; Soygazi and Oğuz, 2025; Yuan and Xi, 2025). Arabic NLP suffers from a shortage of specialized datasets, especially in the medical domain, limiting the ability to train accurate deep learning models (Hammoud et al., 2021; Al-Smadi et al., 2025). Existing datasets are often imbalanced, causing poor performance on minority classes and requiring advanced oversampling techniques (Al-Smadi et al., 2025). The linguistic complexity of Arabic, including morphology, dialects, and orthographic variation, further complicates preprocessing and model training (El Rifai et al., 2022; Al-Smadi, 2024). Additionally, most available corpora are single-label, despite real-world texts requiring multi-label classification, creating a gap in dataset availability and model development (El Rifai et al., 2022; Al-Smadi, 2024; Ehab et al., 2026). Finally, there is a lack of pretrained Arabic biomedical models, making domain adaptation difficult for medical applications (Hammoud et al., 2021; Al-Smadi et al., 2025; El Rifai et al., 2022; Al-Smadi, 2024; Nayel et al., 2023).

Medical text classification is essential because it:

- Supports clinical decision-making by structuring information about symptoms, diagnoses, treatments, and patient intent, improving downstream systems such as adverse event detection and clinical decision support systems (Wang et al., 2023).
- Enhances patient–doctor communication systems, especially intent classification for medical question answering and triage (Yuan and Xi, 2025).
- Enables large scale analysis of electronic health records (EHRs), reducing manual workload and improving efficiency in healthcare data processing (Soygazi and Oğuz, 2025).
- Addresses data scarcity challenges through semi-supervised learning, domain-specific pretrained models (BioBERT, ClinicalBERT, ERNIE Health), and augmentation techniques that improve classification accuracy even with limited labeled data (Wang et al., 2023; Soygazi and Oğuz, 2025).
- Improves robustness and interpretability by integrating external medical knowledge (e.g., knowledge graphs) and advanced attention mechanisms (Yuan and Xi, 2025).

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 394–397
March 28, 2026. ©2026 Association for Computational Linguistics

In this paper, we propose a hierarchical classification model for Arabic medical texts that is designed to capture the fine-grained semantic structures inherent in domain-specific language. In addition, a large language model (LLM) is employed to generate contextualized text representations, leveraging its deep semantic understanding to enhance the text representation.

2 Background

The task is formulated as a single-label, multi-class classification problem under imbalanced data conditions of Arabic medical text. Each instance consists of a healthcare-related question–answer pair provided as a single Arabic text field. The dataset contains 82 medical categories and is highly imbalanced across classes. The category names were originally defined in Arabic and translated into English using an LLM for ease of interpretation, while all input texts remain in Arabic. Given an input text, the systems predict a single integer label corresponding to one of the predefined categories.

The dataset comprises 27,951 labeled training instances and 18,633 blind test instances, with test labels withheld (Gupta et al., 2026).

3 System Overview

The architecture of the proposed system is shown in Figure 1. There are four main phases: Preprocessing, embeddings generation, hierarchal-based training model and evaluation.

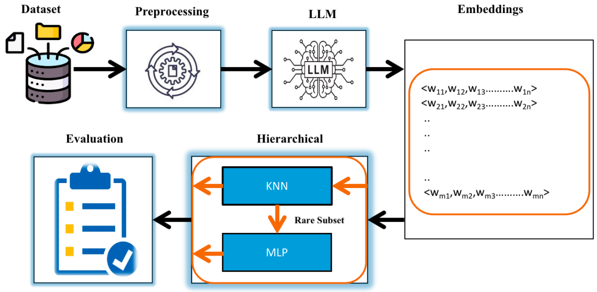
Dataset
Preprocessing
LLM
LLM
Embeddings
$<W_{11},W_{12},W_{13}.........W_{1n}>$
$<W_{21},W_{22},W_{23}.........W_{2n}>$
$<W_{m1},W_{m2},W_{m3}.........W_{mn}>$
Evaluation
Hierarchical
KNN
Rare Subset
MLP

Figure 1: Overall Architecture of the Proposed Model

3.1 Preprocessing

The preprocessing phase consists of various processes including:-

- *Text Cleaning:-* includes removal of stopwrods, diacritics, punctuation marks, symbols, extra white spaces, English letters, and digits.

- *Text Normalization:-* normalize all writing variations to a defined format such as the letters {آ، أ، إ} are normalized to ا, {ى ، ي} are normlized to ى.

- *Balancing Strategy:-* we applied Anchor-Cap balancing strategy using the original dataset as the primary anchor corpus (Mohammadi et al., 2025), applying a cap of 600 samples per class to prevent class dominance. To address class imbalance, we selectively augmented the data using an external Arabic healthcare dataset[1], targeting only underrepresented classes

3.2 Embeddings Generation

We employed ArabicBERT (Safaya et al., 2020) (pretrained BERT-base language model for Arabic) as the base Arabic LLM. The model was fine-tuned for eight epochs using a balanced dataset to adapt it to the target domain. Following fine-tuning, the adapted model was utilized to generate dense vector embeddings for Arabic text. The resulting embeddings were subsequently normalized to ensure compatibility with the cosine similarity–based comparison.

3.3 Hierarchal-Based Model

An initial prediction is generated using a k-Nearest Neighbors (KNN) classifier. If the predicted class corresponds to a rare or underrepresented category, the prediction is subsequently refined using a Multi-Layer Perceptron (MLP) model.

The KNN classifier was configured with 25 neighbors, employing distance-based weighting and the *cosine similarity* metric. The model was trained by fitting it on the complete set of training embeddings, with parallel computation enabled across all available processing cores `n_job = -1`.

Rare labels were identified as those classes with fewer than 50 instances in the whole training set. For these underrepresented classes, a specialized MLP classifier was employed. The training pipeline consisted of standardization using `StandardScaler`, followed by an MLP architecture with hidden layers of sizes 1024–512–256–128, incorporating early stopping and an adaptive learning rate. This model was

[1] https://www.kaggle.com/datasets/mohamedramadan2040/arabic-healthcare/data

trained exclusively on the subset of training data corresponding to rare classes.

3.4 Evaluation

Evaluation of the submitted models is assessed using the macro-averaged F1-score, which is obtained by averaging the F1-scores across all classes.

By assigning equal weight to each class independent of its prevalence, this evaluation criterion emphasizes robust and balanced classification performance, particularly with respect to minority and low-frequency classes.

4 Experimental Setup

In this section, a detailed description of the experimental setup are given, which is important to ensure reproducability of the proposed model. All model configurations, data preprocessing examples, training procedures, and evaluation formula are given to allow other researchers to replicate the experiments and validate the reported results. In this study, all experiments were conducted on the `Google Colab` platform using Python as the primary programming language. Deep learning models were implemented using the **`PyTorch`** library, while classical machine learning models were developed with scikit-learn (**`sklearn`**).

For dataset, a representative dataset instance is presented before and after preprocessing to illustrate the effects of text preprocessing on model-ready data.

Raw instance:-

السؤال

خدر في يدي اليسرى عند ممارسة المشي لمدة طويلة
الجواب

هل الخدر من الكتف للاسفل؟ هل يوجد الالم في الرقبة؟

The instance after preprocessing:-

خدر يدي اليسري عند ممارسه المشي لمده طويله الخدر
الكتف للاسفل ؟ يوجد الالم الرقبه ؟

For the first layer (KNN) in the hierarchical model, Table 1 presents the corresponding parameter settings.

Table 1: Parameter settings for the KNN classifier used in the first layer of the hierarchical model.

Parameter	Value
Number of neighbors ($n\_neighbors$)	25
Weight function ($weights$)	distance
Distance metric ($metric$)	cosine
Number of parallel jobs (n_jobs)	-1

Table 2: Configuration of the MLP classifier pipeline for rare-class specialization.

Component / Parameter	Value
Pipeline Component	StandardScaler MLPClassifier
Hidden layer sizes	(1024, 512, 256, 128)
Activation function	ReLU
Solver	Adam
Regularization term (α)	0.0005
Batch size	64
Learning rate strategy	adaptive
Initial learning rate	0.001
Maximum iterations	100
Random state	42
Early stopping	True
Validation fraction	0.15
Number of iteration	30
Convergence tolerance (tol)	1×10^{-5}
Verbose output	True

5 Results

We have implemented various LLMs on test set, Table 3 reports the results obtained by different models.

Table 3: Comparison of different models on the dataset using macro-averaged F1 score.

Model	Macro F1 Score
ArabicBERT (Safaya et al., 2020)	0.3800
RAC	0.3785
AraBERT	0.3500
CAMeLBERT	0.3047
KNN + MLP (Proposed)	0.4600

6 Conclusion

In this work, we presented a hierarchical classification framework for Arabic text, combining a KNN classifier for general predictions with a specialized MLP model for rare classes. Extensive experiments demonstrated that our approach effectively handles class imbalance and improves performance on underrepresented categories, achieving a macro-averaged F1 score of 0.46, outperforming baseline models including ArabicBERT, RAC, AraBERT, and CAMeLBERT. The results highlight the importance of data quality, preprocessing, and targeted modeling strategies for low-resource and morphologically rich languages like Arabic. Future work will explore more advanced embeddings, multilingual transfer learning, and semi-supervised approaches to further enhance classification performance across all classes.

In this work, we proposed a hierarchical classification system for Arabic text, integrating a KNN-based first layer for general predictions with a specialized MLP model for rare classes. The system leverages fine-tuned embeddings generated from ArabicBERT and incorporates normalization and preprocessing techniques to enhance model performance. Experimental results demonstrate that our approach achieves a macro-averaged F1 score of 0.46, outperforming baseline models such as ArabicBERT Fine-tune, RAC, AraBERT, and CAMeLBERT, particularly in handling the underrepresented classes.

Despite these improvements, the study has some limitations. The reliance on supervised learning restricts generalization to completely unseen or highly imbalanced classes, and the current approach may be sensitive to the quality and size of the training data. Additionally, deeper models or larger embeddings were not explored because of computational constraints.

Future work will focus on addressing these limitations by exploring semi-supervised and self-supervised approaches, multilingual and cross-domain transfer learning, and advanced embedding techniques to further improve performance and robustness. In addition, integrating explainable AI techniques could enhance interpretability, making the system more practical for real-world applications in Arabic text classification.

References

Bushra Al-Smadi, Bassam Hammo, Hossam Faris, and Pedro A. Castillo. 2025. Enhancing the classification of imbalanced arabic medical questions using deepsmote. *AI*, 6(4).

Bushra Salem Al-Smadi. 2024. Deberta-bilstm: A multi-label classification model of arabic medical questions using pre-trained models and deep learning. *Computers in Biology and Medicine*, 170:107921.

Rana Ehab, Ahmed El-Sawy, Mohammed Aldawsari, and Hamada Nayel. 2026. DEAST: A dataset for english-arabic scientific translation and vice versa. *Data in Brief*, 64:112381.

Hozayfa El Rifai, Leen Al Qadi, and Ashraf Elnagar. 2022. Arabic text classification: the need for multi-labeling systems. *Neural Computing and Applications*, 34(2):1135–1159.

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.

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

Hadi Mohammadi, Ehsan Nazerfard, and Mostafa Haghir Chehreghani. 2025. Anchor-based oversampling for imbalanced tabular data via contrastive and adversarial learning.

Hamada Nayel, Nourhan Marzouk, and Ahmed Elsawy. 2023. Named entity recognition for arabic medical texts using deep learning models. In *2023 Intelligent Methods, Systems, and Applications (IMSA)*, pages 281–285.

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.

Fatih Soygazi and Damla Oğuz. 2025. Medical text classification using semisupervised learning and bert-based models. *Mühendislik Bilimleri ve Araştırmaları Dergisi*, 7(1):60–69.

Yu Wang, Yuan Wang, Zhenwan Peng, Feifan Zhang, Luyao Zhou, and Fei Yang. 2023. Medical text classification based on the discriminative pre-training model and prompt-tuning. *DIGITAL HEALTH*, 9:20552076231193213.

Yujia Yuan and Guan Xi. 2025. Msa k-bert: A method for medical text intent classification. *Applied Sciences*, 15(12).