MedArabs at AbjadMed: Arabic Medical Text Classification via Data- and Algorithm-Level Fusion

Amrita Singh
University of New South Wales (UNSW), Sydney

Abstract

In this work, we address the challenges of Arabic medical text classification, focusing on class imbalance and the complexity of the language's morphology. We propose a multi-class classification pipeline based on Data- and Algorithm-Level fusion, which integrates the optimal Back Translation technique for data augmentation with the Class Balanced (CB) loss function to enhance performance. The domain-specific AraBERT model is fine-tuned using this approach, achieving competitive results. On the official test set of the AbjadMed task, our pipeline achieves a Macro-F1 score of 0.4219, and it achieves 0.4068 on the development set.

1 Introduction

The classification of Arabic medical texts has become a critical task in healthcare due to the increasing volume of medical data (Wahdan et al., 2024). With the rapid growth of electronic health records, medical literature, and patient queries, there is a need to organize vast amounts of unstructured information into accessible, meaningful categories (Tayefi et al., 2021). Arabic medical text classification helps identify relevant topics, symptoms, diagnoses, and treatments, enabling healthcare professionals to retrieve pertinent information efficiently and accurately, thereby improving decision-making and patient care (Hammoud et al., 2021). The shared task AbjadMed (Gupta et al., 2026) provides a publicly available dataset consisting of question-answer pairs across multiple medical domains, serving as a benchmark for this work. This task presents unique challenges, particularly due to Arabic's rich morphology and significant class imbalance in the datasets.

We address these challenges by developing a robust classification pipeline through Data- and Algorithm-Level fusion for Arabic medical text classification. In this pipeline, Data- and Algorithm-Level fusion refers to the integration of two key strategies: first, applying optimal data augmentation techniques from the Data-Level, such as back translation, synonym replacement, or random deletion, to enrich the dataset and address issues like class imbalance; and second, optimizing the model's performance with an improved loss function at the Algorithm-Level. By combining these two approaches, we aim to enhance the effectiveness of the domain-specific, pre-trained AraBERT model. The pipeline improves classification performance, tackles class imbalance, and accounts for the language's complexity. Our approach achieves competitive results, ranking in the top-10 entries on the official dev/test set, demonstrating its potential to advance Arabic medical text classification.

2 Related Work

Research on Arabic medical text classification has evolved from small, coarse-grained corpora to large-scale, fine-grained resources, improving performance in healthcare domains (Hammoud et al., 2021). However, challenges like class imbalance and data scarcity are common in fields such as legal (Singh et al., 2025), healthcare (Roy et al., 2024), and software (Goyal, 2025). These issues are often addressed with techniques like data augmentation (ElSabagh et al., 2025), refinement before classification (Singh et al., 2024a), and clustering with hierarchical classification (Singh et al., 2024b). In Arabic medical text classification, these challenges are amplified by the language's rich morphology, absence of diacritics, letter shape variations, and gender agreement, making preprocessing and classification more difficult (Hamzaoui et al., 2025). Additionally, Arabic's diglossia, where Modern Standard Arabic coexists with regional dialects, further complicates tasks, especially in medical texts where formal and informal language mix (Khwaileh et al., 2025). To overcome these challenges, researchers use powerful transformer mod-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 100–104
March 28, 2026. ©2026 Association for Computational Linguistics

els like AraBERT (Antoun et al.), AraGPT-2 (Antoun et al., 2021b), and AraElectra (Antoun et al., 2021a), pre-training them on large Arabic datasets to capture language-specific patterns and improve performance (Wahdan et al., 2024). Techniques to handle class imbalance and enhance classification performance, such as those discussed by Wei and Zou (2019), Sabty et al. (2021), and Abuzayed and Al-Khalifa (2021), have been widely employed. Our approach differs by using a classification pipeline that integrates Data- and Algorithm-Level Fusion for Arabic medical text classification, combining optimal data augmentation with the best loss functions to achieve competitive results.

3 Task and Dataset

Gupta et al. (2026) introduces a shared task that is publicly available on Kaggle[1]. The task is formulated as a multi-class classification problem within the Arabic healthcare domain. The input consists of a question-answer (q&a) pair in Arabic, and the goal is to predict the category that corresponds to the medical domain of the q&a pair. Examples of the task input and output types are shown in Fig. 1. The provided dataset is divided into training, development, and testing sets. The training set consists of $27,951$ q&a pairs across 82 categories. The distribution of the training dataset is depicted in Fig. 2, which highlights the steep class imbalance problem (Henning et al., 2023). The development and testing sets contain $18,634$ q&a pairs, but no labels (categories) are provided. The task is to predict the category for each q&a pair in the development and testing set.

Arabic Medical Text Classification

Input (Q-A pair): السؤال

نتيجه التحليل 10 D عندي نقص في فيتامين

الجواب

يجب مراجعة الطبيب حتى يصف لك المقدار المناسب من مكملات فيتامين د وينصح بالتعرض
للشمس المباشرة مدة 15 دقيقة على الأقل والتركيز على الأطعمة الغنية بهذا الفيتامين مثل منتجات
الألبان، البيض، الأسماك الدهنية كالسلمون والسردين والتونا وزيت السمك

Output (Medical Domain): Vitamins and minerals

Figure 1: Task Overview

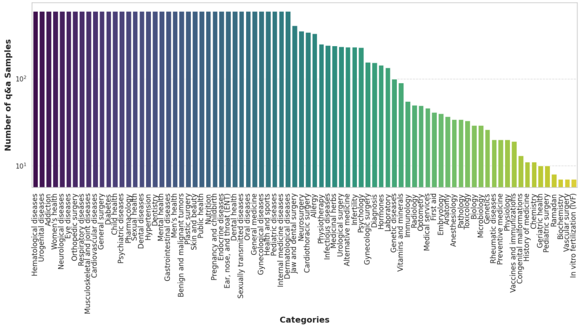

Figure 2: Training Dataset Distribution

4 System Overview

As discussed in Section 3, the training dataset suffers from class imbalance. To address this, we propose a simple yet effective pipeline, as illustrated in Fig. 3. The proposed pipeline consists of three main stages: the Data-Level, the Algorithm-Level, and the final Data- and Algorithm-Level Fusion. These stages are described in the following subsections.

4.1 Data-Level

In the Data-Level, we first split the provided training dataset in an $80 : 20$ ratio, resulting in $22,360$ samples in the sub-training set and $5,591$ samples in the sub-testing set. This internal validation split is used solely for method selection, while the final pipeline is trained on the full training set, as outlined in Section 4.3. We then sample the minority class by extracting categories with fewer than 200 instances from the sub-training set, yielding 32 categories and $1,155$ samples. This represents approximately 5% of the sub-training dataset. Next, we apply various data augmentation techniques, including Back Translation (Arabic-English-Arabic), Synonym Replacement, and Random Deletion, both independently and in combination. We fine-tune the domain-specific AraBERT model and test it on the sub-testing set to identify the optimal data augmentation technique(s). Our experiments show that Back Translation alone outperforms all other techniques and combinations, making it the optimal data augmentation technique.

4.2 Algorithm-Level

In the Algorithm-Level, we split the training dataset in the same manner as done in Section 4.1. We experiment with different loss functions (Henning et al., 2023), including Cross Entropy (CE), Weighted Cross Entropy (WCE), Class Balanced

[1] EACL 2026 Abjad NLP Shared Task 4

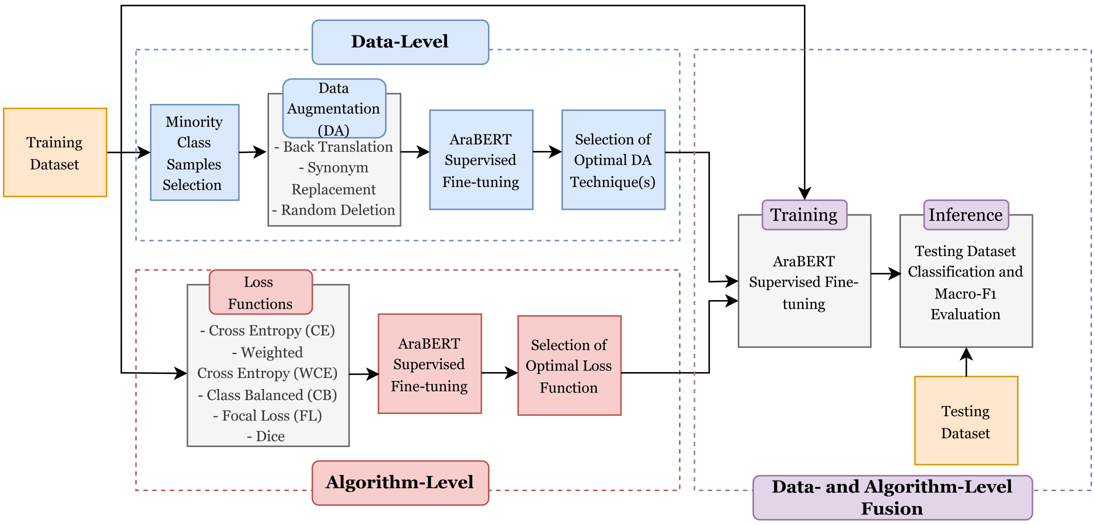
Data-Level
Training Dataset
Minority Class Samples Selection
Data Augmentation (DA)
- Back Translation
- Synonym Replacement
- Random Deletion
AraBERT Supervised Fine-tuning
Selection of Optimal DA Technique(s)
Loss Functions
- Cross Entropy (CE)
- Weighted Cross Entropy (WCE)
- Class Balanced (CB)
- Focal Loss (FL)
- Dice
AraBERT Supervised Fine-tuning
Selection of Optimal Loss Function
Algorithm-Level
Training
AraBERT Supervised Fine-tuning
Inference
Testing Dataset Classification and Macro-F1 Evaluation
Testing Dataset
Data- and Algorithm-Level Fusion

Figure 3: Multi-Class Classification via Data- and Algorithm-Level Fusion

(CB), Dice, and Focal Loss (FL), and fine-tune the AraBERT model. We test the model on the sub-testing set to identify the optimal loss function. Our experiments indicate that the Class Balanced (CB) loss function is the most effective for this task.

4.3 Data- and Algorithm-Level Fusion

In this stage, we perform task-specific fine-tuning using the entire training dataset, which contains $27,951$ samples. We apply the optimal data augmentation technique from the Data-Level, Back Translation (Arabic-English-Arabic), to the minority class samples. We combine this with the optimal loss function from the Algorithm-Level, Class Balanced (CB) loss. The domain-specific AraBERT model is fine-tuned, and its performance is evaluated on the development and testing dataset. The macro-F1 score is calculated to assess model performance.

5 Experimental Setup

For training, we utilize the AbjadMed training dataset, while evaluation is performed on the official development and testing set. Model performance is measured using the Macro-F1 score. The model used is AraBERT, which is fine-tuned for 10 epochs with a learning rate of 2e-5 and a batch size of 16. We experiment with 5 different random seeds on the internal validation split used for method selection (not on the official dev/test set), ranging from 1 to 5, and select the best-performing seed (seed 4) for all subsequent experiments. We implement our pipeline using Hugging Face Transformers (Wolf et al., 2020), and all experiments are conducted on a Google Colab T4 GPU.

6 Result

6.1 Development Set Performance

The results of the proposed pipeline, illustrated in Fig. 3, on the development set are shown in Table 1. Our proposed pipeline achieves a substantially better score on the development set. Due to resource constraints, we test with only a single model, AraBERT, and avoid ensemble techniques, which may further improve performance. Additionally, for diversity, we could also generate samples of minority classes using a domain-specific generation model or GPT-5, but we leave these approaches for future work.

Team	Test Score (Macro-F1)
MedArabs (Ours)	**0.4068**

Table 1: Results on the AbjadMed Subtask official development set, reported in Macro-F1

6.2 Testing Set Performance

Table 2 presents the results of the proposed pipeline on the official test set. Out of 40 submissions, our proposed method ranks in the top-10 leader-board entries.

7 Conclusion

We present a simple yet effective Multi-Class Classification pipeline via Data- and Algorithm-Level Fusion for Arabic medical text classification. Our pipeline combines the optimal data augmentation

Rank	Team	Test Score (Macro-F1)
—	—	—
7	baellouf	0.4398
—	—	–
9	**MedArabs (Ours)**	**0.4219**
—	—	—
19	Kyaw Htin Aung	0.3911
—	—	—
29	Ghader Kurdi (UQU)	0.3393
—	—	—
39	Malaak	0.2460

Table 2: Results on the AbjadMed Subtask official test set: Our proposed pipeline ranks in the top-10, with the reported Macro-F1 score

technique(s) from the Data-Level approach with the best loss function from the Algorithm-Level approach to achieve competitive results. Due to resource constraints, we test with a single model, AraBERT, and leave the exploration of ensemble techniques and alternative data generation models, such as GPT-5 or domain-specific generation models, for future work. Future research will focus on enhancing model performance by incorporating these methods and addressing the class imbalance problem.

References

Abeer Abuzayed and Hend Al-Khalifa. 2021. Sarcasm and sentiment detection in arabic tweets using bert-based models and data augmentation. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 312–317.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021a. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the sixth arabic natural language processing workshop*, pages 191–195.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021b. Aragpt2: Pre-trained transformer for arabic language generation. In *Proceedings of the sixth arabic natural language processing workshop*, pages 196–207.

Ahmed Adel ElSabagh, Shahira Shaaban Azab, and Hesham Ahmed Hefny. 2025. A comprehensive survey on arabic text augmentation: approaches, challenges, and applications. *Neural Computing and Applications*, pages 1–34.

Somya R Goyal. 2025. Current trends in class imbalance learning for software defect prediction. *IEEE Access*.

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 *International Conference on Intelligent Data Engineering and Automated Learning*, pages 196–203. Springer.

Benamar Hamzaoui, Djelloul Bouchiha, and Abdelghani Bouziane. 2025. A comprehensive survey on arabic text classification: progress, challenges, and techniques. *Brazilian Journal of Technology*, 8(1):e77611–e77611.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540.

Tariq Khwaileh, Eiman Mustafawi, Shereen Elbuy, Noor Numan, and Samawiyah Ulde. 2025. Arabic aphasia research through a clinical and linguistic lens: A systematic review of current limitations and future directions. *International Journal of Language & Communication Disorders*, 60(4):e70064.

Debashis Roy, Anandarup Roy, and Utpal Roy. 2024. Learning from imbalanced data in healthcare: State-of-the-art and research challenges. *Computational Intelligence in Healthcare Informatics*, pages 19–32.

Caroline Sabty, Islam Omar, Fady Wasfalla, Mohamed Islam, and Slim Abdennadher. 2021. Data augmentation techniques on arabic data for named entity recognition. *Procedia Computer Science*, 189:292–299.

Amrita Singh, Chirag Jain, Mohit Chaudhary, and Preethu Rose Anish. 2024a. Refining app reviews: Dataset, methodology, and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 595–608.

Amrita Singh, Aditya Joshi, Jiaojiao Jiang, and Hyeyoung Paik. 2025. A survey of classification tasks and approaches for legal contracts. *Artificial Intelligence Review*, 58(12):380.

Amrita Singh, Preethu Rose Anish, Aparna Verma, Sivanthy Venkatesan, Logamurugan V, and Smita Ghaisas. 2024b. A data decomposition-based hierarchical classification method for multi-label classification of contractual obligations for the purpose of their governance. *Scientific Reports*, 14(1):12755.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549.

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).

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.