REIGNITE at AbjadMed: Imbalance-Aware Fine-Tuning of Pretrained Arabic Transformers for Arabic Medical Text Classification Task

Nahid Montasir Rifat*, **Foyez Ahmed Dewan***
Department of Computer Science and Engineering
Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh
{nahidmuntasir2, foyez.ruet767}@gmail.com

Abstract

This paper presents our system developed for the AbjadNLP Shared Task 4 on Medical Text Classification in Arabic, which aims to assign Arabic medical question–answer pairs to a predefined set of medical categories. The task poses significant challenges due to severe class imbalance across 82 categories and the linguistic complexity of domain-specific Arabic medical text. To address these challenges, we propose an imbalance-aware training framework that combines targeted data augmentation for minority classes with class-weighted focal loss during fine-tuning. We evaluate multiple Arabic pretrained transformer models under a unified training configuration and further improve robustness through a majority-voting ensemble of the best-performing models. Our approach achieves competitive performance, ranking **15th** on the private leaderboard with a macro F1 score of **0.4052**, demonstrating the effectiveness of combining different data augmentation techniques, imbalance-aware training objectives, and ensemble learning for large-scale, highly imbalanced Arabic medical text classification. The code is available on GitHub.[1]

1 Introduction

Automatic classification of medical text plays a crucial role in improving access to healthcare information, clinical decision support, and large-scale medical data organization. Transformer-based language models have achieved strong performance across many natural language processing tasks; however, their effectiveness in the medical domain remains challenging, particularly for non-English languages such as Arabic.

A major challenge in real-world Arabic medical text classification datasets is severe class imbalance. This imbalance biases learning toward majority classes and significantly degrades performance on rare categories. Standard training objectives such as cross-entropy loss and naive data sampling strategies are often insufficient in large multi-class settings.

This shared task focuses on Arabic medical text classification and provides a realistic benchmark consisting of question–answer pairs spanning 82 medical categories with extreme variation in class frequencies. To address this imbalance, we employ targeted data augmentation techniques for minority classes, including back-translation, domain-specific synonym replacement, and random word swapping, along with Arabic text normalization. During training, we further mitigate imbalance by incorporating class-weighted focal loss.

The main contributions of this work are summarized as follows:

- A transformer-based framework for Arabic medical question–answer classification under severe class imbalance.
- Integration of targeted data augmentation with class-weighted focal loss to mitigate label imbalance during training.
- A majority-voting ensemble of three fine-tuned Arabic transformer models to improve generalization across classes.

2 Related Work

Transformer-based models have shown strong performance across text classification tasks in multilingual and low-resource settings. In multilingual classification, XLM-RoBERTa has been applied to hate speech detection in English, German, and Hindi, achieving high macro F1 scores (Roy et al., 2021). These results highlight multilingual pretraining for linguistic diversity, though such studies

*Equal contribution.

[1] https://github.com/NahidMuntasir7/AbjadNLP-at-EACL-2026

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 132–136

usually focus on social media rather than medical text.

For Arabic, several studies demonstrate that Arabic-specific pretrained transformers outperform traditional deep learning models. Antoun et al. (Antoun et al., 2020) introduced AraBERT, achieving state-of-the-art results in sentiment analysis, named entity recognition, and question answering. Later work applied AraBERT to domain-specific tasks, such as telecom customer satisfaction analysis, consistently outperforming CNN and RNN baselines (Aftan and Shah, 2023), emphasizing language- and domain-aware pretraining.

Recent studies explore alternative Arabic transformer architectures. CAMeLBERT outperformed PaLM for Arabic sentiment and entity-level classification despite smaller size (Almutrash and Abudalfa, 2024). MARBERT, pretrained on dialectal and social media text, performed well in multi-label sentiment and spam detection (Alotaibi et al., 2022). These models capture complementary linguistic characteristics, motivating comparative evaluation and ensemble strategies.

Data scarcity and class imbalance remain key challenges. Feng et al. (Feng et al., 2021) surveyed data augmentation techniques, showing that combining back-translation, synonym replacement, and word-level perturbations improves robustness. Transformer ensembles also outperform single models in tasks like sentiment analysis and topic modeling (Zhang and Shafiq, 2024). However, prior work rarely addresses extreme class imbalance in Arabic medical text. Our work bridges this gap by integrating targeted augmentation, imbalance-aware training, and transformer ensembles in a unified framework.

3 Task Overview

This shared task (Gupta et al., 2026) focuses on categorizing Arabic medical question–answer pairs into a predefined set of medical labels.

3.1 Dataset Description

The dataset (Nagarajan et al., 2025) is released with predefined training and test splits. The training set contains 27,951 samples spanning 82 categories, while the test set consists of 18,634 samples. The dataset exhibits severe class imbalance, with category frequencies ranging from as few as 7 samples to over 600 samples per class. Figure 1 visualizes the class imbalance by sorting classes by their number of samples and highlighting underrepresented, below-average, and above-average categories. It is evident that nearly one-third of the categories contain fewer than 100 training samples, highlighting the challenging imbalance characteristics of the dataset.

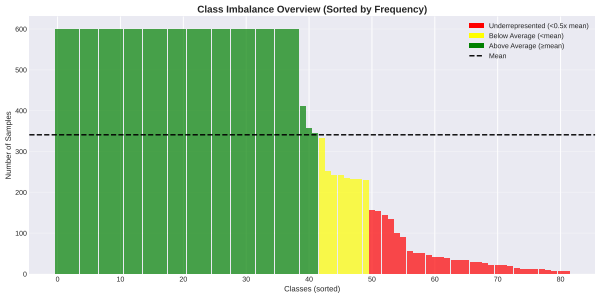

Figure 1: Class imbalance in the training set, sorted by sample count, with colors indicating relative class frequency and a dashed line showing the mean.

4 System Description

4.1 Text Preprocessing

To ensure high-quality input for transformer-based models, several preprocessing steps were applied to the Arabic medical text. These included the removal of Arabic diacritics, normalization of character variants such as Alif, Ya, and Ta Marbuta, and elimination of extra whitespace. These normalization steps reduce orthographic variability while preserving the semantic content of each question–answer pair.

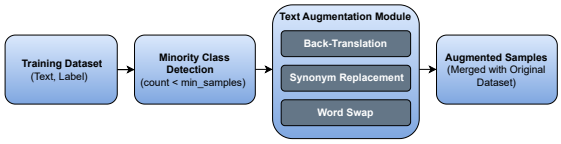

Figure 2: Overview of the data augmentation strategy for minority classes.

4.2 Data Augmentation Techniques

To address extreme class imbalance, targeted data augmentation was applied to minority classes, defined as categories with fewer than 50 training samples. Each minority class was augmented to reach a target size of 150 instances by randomly sampling original examples with replacement and applying three complementary augmentation strategies until the target size was achieved. Approximately 60% of the augmented samples were generated via **back-**

translation, using Arabic–English and English–Arabic models from the Helsinki-NLP OPUS-MT collection to create paraphrases. Around 30% were generated via **synonym replacement**, leveraging a manually constructed domain-specific Arabic medical synonym lexicon to replace up to three words per sentence, and the remaining 10% were produced using **random word swaps**, exchanging one to two word positions in sufficiently long sentences to introduce minor structural variation while preserving semantics. All sampling and augmentation procedures were conducted using a fixed random seed of 42 to ensure reproducibility. The augmentation pipeline is illustrated in Figure 2.

4.3 Training Pipeline and Class Imbalance Handling

The preprocessed and augmented data were split into 80% training and 20% validation sets using stratified sampling to preserve class distributions. Multiple transformer-based models were fine-tuned, including MARBERTv2, XLM-RoBERTa (base), CAMeLBERT, AraBERT-base v2, AraBERT-base v02 and AraBERT-large v02. To ensure unbiased comparisons, all models were trained using identical hyperparameter configurations, as detailed in Table 1. Early stopping was employed to prevent overfitting, and model evaluation performed at the end of each epoch. The best-performing checkpoint, determined by validation macro- F1 score, was retained for final evaluation.

Table 1: Hyperparameter configuration used for all transformer-based models.

Hyperparameter	Value
Maximum sequence length	384
Batch size	16
Learning rate	2e-5
Number of epochs	15
Warmup steps	500
Weight decay	0.01
Random seed	42
Early stopping patience	3

To mitigate the effect of class imbalance, a **class-weighted focal loss** was employed during training. Class weights were computed using an inverse-frequency scheme and clipped to the range [0.5, 10.0] to prevent instability from excessively high weights in extremely rare classes. The focal loss applied a focusing parameter $\gamma = 2.5$, down-weighting well-classified examples and emphasizing hard or misclassified instances, thereby allowing the model to prioritize learning from rare and difficult examples. Loss computation was performed on a per-sample basis and aggregated using mean reduction, ensuring stable optimization while preserving the effects of both class weighting and focal modulation.

4.4 Model Ensemble

After fine-tuning, predictions were generated on the test set using the best-performing models, specifically CAMeLBERT, AraBERT-base v02, and AraBERT-large v02. The final prediction was determined using a **majority voting ensemble** of these three models. The overall workflow of the system, including preprocessing, data augmentation, model training, and ensemble, is depicted in Figure 3.

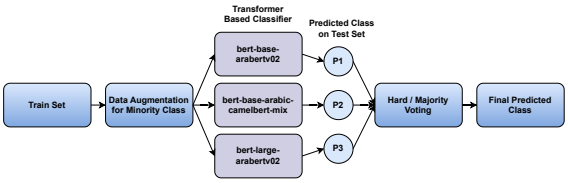

Figure 3: End-to-end workflow of the proposed system.

5 Results

5.1 Initial Model Performance

The baseline performance of the transformer-based models trained on the original dataset is summarized in Table 2. Among the individual models trained on the original dataset, AraBERT-base v02 achieved the highest macro F1 score on the public validation set (0.3718), while AraBERT-large v02 obtained the best performance on the private validation set with a macro F1 score of 0.3974. CAMeL BERT followed closely with macro F1 scores of 0.3701 and 0.3856 on the public and private validation sets, respectively. Other models such as MARBERTv2, XLM-RoBERTa, and AraBERT-base v2 achieved lower performance, highlighting the variability in representational power among different transformer architectures when applied to highly imbalanced medical text data.

5.2 Performance After Data Augmentation

Incorporating targeted data augmentation for minority classes led to consistent improvements

Table 2: Macro F1 scores on public and private validation sets for individual models, augmented models, and ensemble.

Training Data	Model Name	Macro F1 (public)	Macro F1 (private)
Original	MARBERTv2	0.3411	0.3652
Original	XLM-RoBERTa	0.3284	0.3422
Original	CAMeLBERT	0.3701	0.3856
Original	AraBERT-base v2	0.3344	0.3517
Original	AraBERT-base v02	0.3718	0.3836
Original	AraBERT-large v02	0.3653	0.3974
Original + Augmented	CAMeLBERT	0.3759	0.3728
Original + Augmented	AraBERT-base v02	0.3816	0.3925
Original + Augmented	AraBERT-large v02	0.3780	0.3858
Ensemble (Original + Augmented)	CAMeLBERT + AraBERT-base v02 + AraBERT-large v02	**0.4048**	**0.4052**

across all models. Table 2 shows that CAMeLBERT, AraBERT-base v02, and AraBERT-large v02, when trained on the augmented dataset, achieved macro F1 scores of 0.3759, 0.3816, and 0.3780 on the public validation set, respectively. This corresponds to relative improvements of 0.6–1.3 points over their counterparts trained on the original dataset. On the private validation set, augmentation improved performance for AraBERT-base v02, increasing macro F1 from 0.3836 to 0.3925. This corresponds to modest improvements of 0.6–1.3 Macro-F1 points per model, indicating quantifiable gains for minority classes.

5.3 Ensemble Performance

To further improve performance, we employed a majority voting ensemble comprising CAMeLBERT and AraBERT-base v02 and AraBERT-large v02, each of which was fine-tuned on the original and the augmented dataset as well. From Table 2, the majority voting ensemble achieved a macro F1 of 0.4048 and 0.4052 on the public and private validation sets, respectively, slightly higher than the best individual model. This suggests the ensemble provides small but consistent gains, particularly for minority classes, as reflected in the Macro-F1 scores.

6 Conclusion

In this work, we addressed Arabic medical text classification under extreme class imbalance. We demonstrated that careful preprocessing, coupled with targeted data augmentation and a class-weighted focal loss, can modestly improve model performance, particularly when combined with ensembling of multiple transformer models. We captured complementary semantic representations by fine-tuning several transformer-based models including CAMeLBERT and various AraBERT variants. Finally, a majority-voting ensemble of the top-performing models further enhanced robustness and achieved the highest macro F1 scores on both public and private validation sets. Our results highlight the effectiveness of combining augmentation, class imbalance handling, and ensembling for improving performance on challenging, highly imbalanced medical text datasets.

7 Limitations

Our studies were limited by the availability of GPUs, which influenced the extent of fine-tuning and unfreezing of the model's parameters. For example, while fine-tuning AraBERT-large v02, only the last 18 layers of the model were unfrozen, which corresponds to approximately 60% of the total parameters due to memory constraints. Further, the time taken for both training and evaluation influenced the level of data augmentation that could be carried out. Although there were classes with up to 600 instances, the number of samples for the minority classes was increased from the original number to 150 to reduce the influence of the imbalance on the model's performance and limit the time taken for the model to train and evaluate. This implies that the models may not perform well on the rare and highly imbalanced classes and that future studies may consider more extensive data augmentation and model scaling to improve the model's performance on the minority classes. We note that our experiments do not include a full component-level ablation; therefore, the reported gains should be interpreted as the combined effect of augmentation, focal loss, and ensembling rather than isolated contributions.

References

Sulaiman Aftan and Habib Shah. 2023. Using the arabert model for customer satisfaction classification of telecom sectors in saudi arabia. *Brain Sciences*, 13(1).

Salman Almutrash and Shadi Abudalfa. 2024. Comparative study on the efficiency of using palm and camelbert for arabic entity sentiment classification. Publisher Copyright: © 2024 1st Saudi Conference on Information Systems, SaudiCIS 2024. All rights reserved.; 1st Saudi Conference on Information Systems, SaudiCIS 2024 ; Conference date: 19-11-2024 Through 21-11-2024.

Abrar Alotaibi, Atta ur Rahman, Raheel Alhaza, Wala Alkhalifa, Narjes Alhajjaj, Atheer Alharthi, Dhai Abushoumi, Maryam Alqahtani, and Dania Alkhulaifi. 2022. Spam and sentiment detection in arabic tweets using marbert model. *Mathematical Modelling of Engineering Problems*.

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, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

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.

Balaji Nagarajan, Niranjan kumar M, and Pranav Gupta. 2025. Eacl 2026 abjad nlp shared task 4. `https://kaggle.com/competitions/eacl-2026-abjad-nlp-shared-task-medical-text-classification-in-arabic`. Kaggle.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. *CoRR*, abs/2101.03207.

Hongzhi Zhang and M Omair Shafiq. 2024. Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of big Data*, 11(1):25.