CasbAI at AraSentEval Shared Task: Robust Dialectal Arabic Sentiment Classification via Multi-Seed Ensembling and Data Augmentation.

Chaima Abdelaziz[1], **KahinaHouda Saadaoui**[1], **Faiza Belbachir**[2], **Lynda Said Lhadj**[1]
[1] Higher National School of Computer Science ESI (Ex. INI), Algiers, Algeria
[2] LyRIDS, ECE Research Center, 75015 Paris, France
{lc_abdelaziz, kk_saadaoui, l_said_lhadj}@esi.dz, fbelbachir@ece.fr

Abstract

This paper describes the system we designed for our participation in the AraSentEval 2026 shared task on Arabic dialectal sentiment analysis. We propose a transformer-based approach relying on MARBERT combined with a multi-seed ensemble strategy and several optimization techniques. Our system integrates seven independently trained models with different random initializations and applies Stochastic Weight Averaging (SWA) to improve generalization. To address class imbalance, we augment the training data through dialectal synonym replacement, increasing the dataset size by 13.9% while preserving dialect distribution. In addition, we incorporate Test-Time Augmentation (TTA) and investigate the use of pseudo-labeling based on high-confidence predictions. We report our experiments on the official dataset covering Moroccan, Egyptian, Jordanian, and Saudi dialects, and analyze the contribution of each component through ablation experiments. Our system achieved a macro F1-score of 84.62% on the test set, ranking 3rd among 15 participating teams.

Keywords: Arabic Sentiment Analysis, Dialectal Arabic, Data Augmentation, Stochastic Weight Averaging

1. Introduction

Sentiment analysis enables automatic detection of opinions in user-generated content and remains a central NLP task. The AraSentEval 2026 shared task (Ezzini et al., 2026), organised within the OSACT7 Workshop at LREC 2026, advances research in Arabic sentiment understanding. We participate in Subtask 1, focusing on multi-class sentiment classification for Arabic dialectal texts.

Arabic dialects dominate social media yet remain under-resourced compared to Modern Standard Arabic (MSA). Their wide variation in phonology, morphology, and orthography compounded by code-switching, non-standard spelling, and emojis, creates significant classification challenges. Transformer-based models such as MARBERT (Abdul-Mageed et al., 2021), pre-trained on 128M Arabic tweets, have shown strong dialectal performance, but small datasets increase sensitivity to initialisation and overfitting. To address this, we propose a system combining three complementary strategies: **multi-seed ensembling** (7 models) to reduce variance, **Stochastic Weight Averaging (SWA)** to improve generalisation, and **Test-Time Augmentation (TTA)** with dialectal synonym replacements to enhance prediction robustness.

Our system achieves a macro F1 of 84.62% on the official test set, ranking 3rd out of 15 teams, confirming the suitability of dialect-specialised pre-trained models for robust Arabic sentiment analysis. Code and notebooks are publicly available on our GitHub repository. The remainder of this paper is organised as follows: Section 2 reviews related work and the dataset; Section 3 describes the system architecture; Section 4 presents results and analysis; Section 5 concludes.

2. Background and Related Work

2.1. Related Work

Arabic sentiment analysis has evolved from lexicon-based and classical machine learning approaches (Badaro et al., 2014; Mohammad et al., 2016; Mikolov et al., 2013) to deep neural architectures (Kalchbrenner et al., 2014; Yin, 2024), with Transformer-based models now establishing the dominant paradigm.

The introduction of BERT (Devlin et al., 2019) enabled the development of Arabic-specific pre-trained models. AraBERT (Antoun et al., 2021) established strong baselines for Modern Standard Arabic (MSA), while CAMeLBERT (Inoue et al., 2021) demonstrated that combining MSA and dialectal pre-training data improves robustness across downstream tasks. Dialect-focused models further highlighted the importance of variant-aware pre-training: DziriBERT (Abdaoui et al., 2022) targeted Algerian Arabic, and MARBERT (Abdul-Mageed et al., 2021), trained on one billion Arabic tweets covering multiple dialects, showed strong performance on informal and social media text.

More recently, large language models (LLMs) have been applied to Arabic sentiment analysis. Studies benchmarking Qwen2.5, DeepSeek-R1, and LLaMA-3 (Alharbi et al., 2025a) re-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 278–283
May 11, 2026.

port competitive zero-shot and in-context learning performance across MSA and several dialects. Instruction-tuned Arabic LLMs such as AceGPT (Huang et al., 2024) and Jais (Sengupta et al., 2023) improve alignment with Arabic linguistic norms. However, LLM-based approaches remain computationally demanding and may exhibit performance variability on low-resource dialects.

These findings motivate our approach: combining dialect-specialized pre-training with robust ensemble strategies and targeted data augmentation.

2.2. AraSentEval SharedTask and Corpus

The AraSentEval 2026 Subtask 1 dataset (Alharbi et al., 2025c) consists of Arabic hotel reviews designed for three-class sentiment classification: positive, negative, and neutral. The dataset was originally collected in Modern Standard Arabic (MSA) and subsequently extended to dialectal varieties through a controlled translation pipeline. MSA reviews were translated into Saudi dialect and Moroccan Darija using a large neural machine translation model (Meta's NLLB-200). To ensure dialectal authenticity and sentiment consistency, all translations were manually validated and corrected by native speakers.

The organizers provide two official splits: A training set and a Test set. The training set is labeled and used for model learning and validation, while the test set is reserved for official evaluation and its labels are not publicly available. The corpus contains user-generated content reflecting multiple Arabic dialects, including Moroccan (Darija), Saudi, Jordanian, and Egyptian. The informal nature of social media text introduces non-standard spelling, dialectal variation, which increase classification complexity.

Corpus	Positive	Negative	Neutral	Total
Train	609	657	465	1731
Test	-	-	-	312

Table 1: Distribution of Sentences per Sentiment Class

The class distribution in the training data is shown in Table 2, illustrating a balanced representation of Arabic dialects.

Dialect	**Sentences**	**%**
Moroccan	433	25.04%
Saudi	433	25.04%
Jordanian	433	25.04%
Egyptian	432	24.96%

Table 2: Dialect distribution in the training set.

This balanced distribution reduces strong bias toward a single sentiment category while preserving realistic sentiment variation across dialects. AraSentEval provides a structured benchmark for comparing sentiment analysis systems across Arabic dialects, building on the previous Ahasis shared task (Alharbi et al., 2025b). This edition attracted 15 participating teams, reflecting growing interest in dialect-aware modelling. Systems are evaluated primarily by Macro F1-score to account for class imbalance, with accuracy, precision, recall, and per-class F1-score as secondary metrics. (Alharbi et al., 2025c).

3. System Description

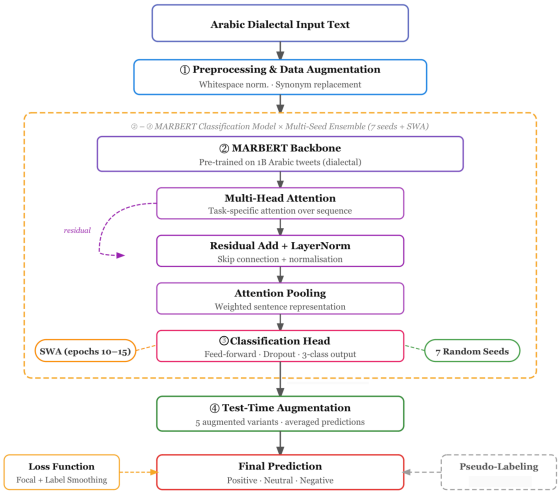

Figure 1: Overall architecture of our system.

Our system is a modular pipeline combining a transformer-based encoder with several robustness-enhancing strategies. As illustrated in Figure 1, it consists of four main components: (1) preprocessing and data augmentation, (2) a MARBERT-based classification model, (3) a multi-seed ensemble with Stochastic Weight Averaging (SWA), and (4) Test-Time Augmentation (TTA) at inference.

3.1. Preprocessing and Data Augmentation

To preserve dialectal characteristics, we apply only minimal preprocessing (whitespace normalization). The original training set contains 1,731 samples across three sentiment classes (657 negative, 609 positive, 465 neutral). To mitigate class imbalance and improve model robustness, we employ a **targeted data augmentation strategy** based on light, stochastic synonym replacement using a small, curated dialectal lexicon:

رائع (*wonderful*) → حلو، جميل، ممتاز (*excellent, beautiful, nice*)

جيد (*good*) → كويس، تمام، منيح (*okay, fine, decent*)
ممتاز (*excellent*) → رائع، جميل (*wonderful, beautiful*)
سيء (*bad*) → خايب، وحش (*poor, awful*)
مشكلة (*problem*) → إشكال، قضية (*issue, matter*)
كان (*was*) → كانت، صار (*was/fem., became*)
الفندق (*the hotel*) → الأوتيل، المكان (*the hotel/dial., the place*)

For each sentence, a single word is randomly replaced with a probability of 0.5, making the procedure **stochastic** and avoiding exact duplicates. This simple yet effective augmentation increases dataset diversity, balances class distributions, and exposes the model to dialectal lexical variations, which are crucial for generalization on small and heterogeneous datasets. After augmentation, all classes reach 657 samples, resulting in a balanced training set of 1,971 examples (+13.9%).

3.2. MARBERT Classification Model

In preliminary experiments comparing several Arabic pre-trained models MARBERT ,BERT-base transformer,(Abdalaa et al., 2021) achieved the highest validation accuracy (94.47%), outperforming AraBERTv2 (81.87%) (Antoun et al., 2021), DziriBERT (84.21%) (Abdaoui et al., 2022), CAMeLBERT-Mix (86.87%), and CAMeLBERT-MSA (87.84%) (Inoue et al., 2021), consistent with prior work on dialectal Arabic sentiment analysis (Belbachir, 2023).

We extend MARBERT with three lightweight task-specific layers. First, a **multi-head attention layer** is added on top of the encoder output to capture task-specific dependencies across the sequence. Its output is integrated via a **residual connection and layer normalization**, which stabilises training and ensures unobstructed gradient flow through the pre-trained layers. Rather than relying solely on the `[CLS]` token, we employ **attention-based pooling** over all token representations, assigning learned importance weights to each position. This is particularly beneficial in dialectal Arabic, where negation markers can appear anywhere and critically determine sentence polarity . For instance, removing مو (*mou*, negation marker meaning *"not"*) from الخدمة مو زينة ("The service is not good") fully reverses the sentiment to الخدمة زينة ("The service is good"). Finally, the pooled representation is fed into a **two-layer classification head** (Linear–LayerNorm–GELU–Dropout(0.35)–Linear) that outputs three logits over the sentiment classes. The full model comprises approximately 168M parameters (165M from MARBERT, 3M added).

3.3. Ensemble and Inference Strategy

To reduce sensitivity to random initialisation, we train seven independent models with seeds $\{42, 123, 2024, 777, 1337, 9999, 555\}$ and average their softmax probability vectors as the ensemble prediction. For each seed, we apply **Stochastic Weight Averaging (SWA)** starting at epoch 10, averaging model weights over the final 5 epochs at a learning rate of 4×10^{-6}. SWA encourages convergence to flatter loss minima, improving generalisation. The per-seed prediction combines 40% of the regular model output and 60% of the SWA output, and these are then averaged across all seven seeds. At inference, we additionally apply **Test-Time Augmentation (TTA)**: each test sample is expanded into five variants through the same synonym replacement procedure used during training, and the final prediction is the mean over all augmented versions. TTA reduces sensitivity to incidental surface variations, complementing the diversity introduced by the multi-seed ensemble.

4. Experimental Setup

We fine-tune MARBERT with a maximum sequence length of 128 tokens, batch size 16 with gradient accumulation over 2 steps, and the AdamW optimiser (β_1=0.9, β_2=0.999, ε=10^{-8}, weight decay 0.01). We use a learning rate of 1.8×10^{-5} with a cosine schedule and 10% linear warmup, and dropout 0.35. All models are trained for 15 epochs with mixed-precision (AMP) on an NVIDIA T4 GPU ($\sim$4.5 min per run).

Loss function. We combine Focal Loss and Label Smoothing Cross-Entropy in a fixed-weight sum:

$$\mathcal{L} = 0.7\,\mathcal{L}_{focal} + 0.3\,\mathcal{L}_{smooth}$$

Focal Loss (γ=2.5) down-weights easy examples and focuses training on harder instances, while label smoothing (ε=0.1) prevents overconfident predictions. Class-frequency inverse weights are applied to both terms to further address the residual imbalance.

5. Results and Analysis

Our system achieved 84.62% macro F1-score, securing 3rd place. Furthermore, Table 3 highlights our performance in comparison to the four other teams.

5.1. Ablation Study

To better understand the contribution of each component, we conduct an ablation study on a 10% stratified validation split of the training data.

Rank	Team	Macro F1	Accuracy
1	musj1984	94.23%	94.23%
2	alighabusaleh	92.63%	92.63%
3	**Ours**	**84.62**%	**84.62**%
4	UoTripoli	84.29%	84.29%
5	elham42	83.33%	83.33%

Table 3: Performance comparison and ranking of the top five systems on the test set

Table 4 reports the macro F1-score obtained by incrementally adding each component to the MARBERT baseline.

Configuration	**Validation F1**
MARBERT baseline (1 seed)	0.9447
+ Light augmentation	0.9509 (+0.61%)
+ Multi-seed ensemble (7)	0.9560 (+0.52%)
+ SWA	0.9619 (+0.59%)
+ TTA	0.9625 (+0.06%)

Table 4: Ablation study

The ablation results indicate that each component contributes positively to the overall performance. Light augmentation provides the largest single improvement (+0.61%), highlighting the importance of increasing lexical diversity in low-resource dialectal settings. SWA adds a further +0.59% by encouraging convergence toward flatter minima and improving generalization. The multi-seed ensemble contributes +0.52%, reducing variance due to random initialization. Finally, TTA brings a modest but consistent +0.06% improvement at no additional training cost. Overall, the combination of all components yields a cumulative gain of +1.78% over the single-seed baseline.

5.2. Prediction Confidence

The ensemble produces well-calibrated predictions, with a median confidence of 93.2% and only 8 samples (2.6%) below 60%. However, no prediction exceeded the 95% pseudo-labeling threshold, suggesting that the model does not reach the certainty level needed to self-train on unlabelled data. The predicted sentiment distribution shows a slight positive bias (40.1% positive, 34.6% negative, 25.3% neutral), consistent with the positive skew typical of hotel review corpora.

5.3. Error Analysis

We conduct an error analysis on a 10% stratified validation split (198 samples) using a single model (seed 42), as ensemble averaging tends to suppress borderline predictions and limit qualitative interpretation. The model achieves a macro F1 of 93.42%, with only 13 errors (6.6%). As shown in Table 5, the *negative* class achieves the highest precision (0.983) but lowest recall (0.879), indicating a tendency to absorb negative samples into the neutral class, which acts as an attractor with the highest recall (0.970).

Class	**P**	**R**	**F1**	**Sup.**
Negative	0.983	0.879	0.928	66
Neutral	0.901	0.970	0.934	66
Positive	0.926	0.955	0.940	66
Macro	0.937	0.934	0.934	198

Table 5: Per-class metrics on the validation set.

Errors concentrate around the neutral frontier: 4 negative samples are predicted as neutral and 4 as positive , while no positive sample is ever predicted as negative, confirming that the model handles extreme polarity boundaries well (see confusion matrix in Figure 2 in Appendix A). Dialect-wise, Darija yields the highest error rate (9.0%), followed by Egyptian (8.2%), Jordanian (5.3%), and Saudi (2.3%). The dominant failure mode in Darija is the over-generalisation of positive lexical cues under conditional or sarcastic framing, and the mishandling of concessive negation, where the model over-weights a negative surface marker while ignoring a subsequent clause that partially redeems the sentiment. Representative examples are provided in Appendix A.

6. Conclusion

In this paper, we presented our system for Arabic dialectal sentiment analysis in AraSentEval 2026. Our approach combines MARBERT fine-tuning with multi-seed training, Stochastic Weight Averaging (SWA), and Test-Time Augmentation (TTA), achieving 84.62% macro F1-score and ranking 3rd among participating teams. The ablation study showed that data augmentation provides the largest improvement, highlighting its importance when training on small datasets.

However, several limitations remain. First, the training set is relatively small, which limits the model's ability to fully capture dialectal variation. Second, the synonym-based augmentation relies on a small manually created lexicon and does not consider context, which may introduce noise. Third, the ensemble approach requires training and storing multiple models, increasing computational cost.

Future work will focus on expanding data augmentation with more context-aware methods, exploring semi-supervised learning on additional unlabeled dialectal data, and investigating more efficient ensemble or parameter-efficient fine-tuning strategies. Modeling dialect information explicitly may also help improve performance across different Arabic varieties.

7. Acknowledgements

We thank the AraSentEval 2026 organizers for providing the shared task, and the anonymous reviewers for their valuable feedback.

References

Mustafa Abdalaa, Richard Sutcliffe, Xia Sun, Jun Feng, Eiad Almekhlafi, and Ephrem Retta. 2021. Improving arabic sentiment analysis using cnn-based architectures and text preprocessing. *Computational Intelligence and Neuroscience*.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2022. Dziribert: a pre-trained language model for the algerian dialect.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Marbert: A robust pre-trained model for arabic social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 7088–7105.

Maram Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025a. Evaluating large language models on arabic dialect sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Maram I Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025b. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 1–6.

Maram I Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025c. Ador: Dataset for arabic dialects in hotel reviews: A human benchmark for sentiment analysis. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 187–191.

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

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173.

Faiza Belbachir. 2023. Foul at semeval-2023 task 12: Marbert language model and lexical filtering for sentiment analysis of tweets in algerian arabic. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval 2023)*, pages 389–396.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*.

Saad Ezzini, Shadi Abudalfa, Maram Alharbi, Salmane Chafik, Hamzah Luqman, Mo El-Haj, Paul Rayson, Samaher Alghamdi, and Reem Alotaibi. 2026. AraSentEval: A shared task on sentiment analysis and swapping in arabic. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Huang Huang, Fei Yu, Jianqing Zhu, Xuming Sun, Qi Zhang, Jizhong Wang, and Fei Huang. 2024. AceGPT: Localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 6383–6396, Mexico City, Mexico.

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.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. Sentiment lexicons for arabic social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Xu, Haewon Kim, Hitesh Bhatt, et al. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. In *arXiv preprint arXiv:2308.16149*.

Qixuan Yin. 2024. Three-class text sentiment analysis based on lstm. *arXiv preprint arXiv:2412.17347*.

8. Appendix

A. Error Analysis - Examples

Lexical shortcut in Darija.

فندق آشمور هاوس كايكون مزيان بزاف من الرفاهية
(*Ashmore House hotel would be very nice in terms of luxury.*)
True: *negative* — Predicted: *positive* — conf=0.935

Temporal sentiment drift.

الفندق كان ممتاز قبل كده بس بقى مستواه أقل عن السنين اللي فاتت
(*The hotel used to be excellent but its level has declined.*)
True: *positive* — Predicted: *neutral* — conf=0.912

Concessive negation.

رأس الدش ما كانش شغال كويس [...] بس الحاجة الإيجابية...
(*The shower was not working well [...] but the positive thing was...*)
True: *neutral* — Predicted: *negative* — conf=0.912

The negation marker ما كانش dominates the prediction while the concessive clause بس (*but*) is insufficiently weighted.

B. Error Analysis - confusion matrix

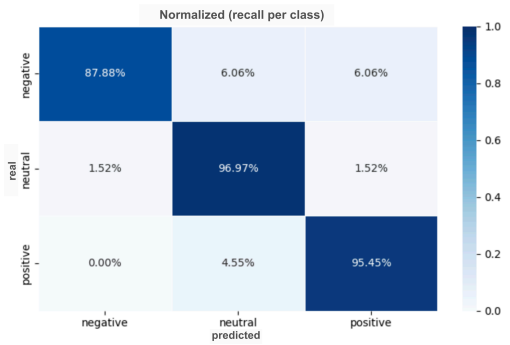

Figure 2: Confusion matrix on the validation set.