BDSI at AraSentEval Shared Task : A Multi-Transformer Contrastive Learning for Arabic Dialect Sentiment Analysis

Kaouthar Elyoussoufi, Mohamed M'haouach, Abdessamad Benlahbib, Hamza Alami
L3IA Laboratory, Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, 30003, Morocco
{kaouthar.elyoussoufi, mohamed.mhaouach, abdessamad.benlahbib,
hamza.alami5}@usmba.ac.ma

Abstract

This paper presents our system for the AraSentEval 2026 shared task on Arabic dialect sentiment analysis. We propose a multi-model ensemble combining AraBERTv2 and CAMeLBERT with supervised contrastive learning to improve sentiment classification. The system incorporates dialect-aware preprocessing, class-weighted cross-entropy loss with label smoothing, supervised contrastive loss for enhanced sentence representations, and rule-based post-processing for dialect-specific patterns. Our approach achieves a macro F1-score of 0.83 on the official test set, demonstrating the effectiveness of contrastive learning with pretrained Arabic language models for dialectal sentiment analysis.

Keywords: Arabic Dialects, Sentiment Analysis, AraBERT, CAMeLBERT, Contrastive Learning, Ensemble Methods

1. Introduction

The rapid growth of user-generated content, particularly in the hospitality domain, has made sentiment analysis an important research area. In Arabic, this content is typically written in a mixture of Modern Standard Arabic (MSA) and diverse regional dialects, which differ significantly in their lexical, morphological, and syntactic properties. The lack of standardized orthography and the limited availability of annotated resources further increase the difficulty of developing robust sentiment analysis systems for Arabic dialects.

The AraSentEval 2026 (Ezzini et al., 2026; Alharbi et al., 2025c,a,b) shared task addresses these challenges by providing a unified benchmark for sentiment analysis across multiple Arabic dialects. Subtask 1 focuses on multi-class Arabic Dialect Sentiment Analysis, requiring systems to classify text into positive, negative, or neutral categories, while remaining robust to linguistic variability.

In this paper, we present our system for Subtask 1 of AraSentEval 2026. Our approach relies on an ensemble of pretrained Arabic transformer models, AraBERTv2 and CAMeLBERT, chosen for their complementary strengths in modeling MSA and dialectal Arabic. To enhance representation learning and class discrimination, we integrate supervised contrastive learning into the training process. In addition, dialect-aware preprocessing and weighted loss functions are employed to address orthographic variation and class imbalance.

The remainder of this paper is organized as follows: Section 2 details the proposed methodology, Section 3 presents experimental results, and Section 4 concludes the paper.

2. Methodology

Our system for Arabic dialect sentiment analysis employs a **dual-model ensemble** architecture combining **AraBERTv2** (Antoun et al., 2020) and **CAMeLBERT** (Inoue et al., 2021), enhanced with **supervised contrastive learning** (Khosla et al., 2020) and dialect-aware post-processing. Each base model is fine-tuned using a robust training strategy designed to address the linguistic variability and class imbalance inherent in dialectal Arabic data.

2.1. Preprocessing Pipeline

We adopt a model-specific preprocessing strategy to maximize compatibility with each transformer's pre-training distribution:

1. **Label Encoding**: Sentiment labels (`positive`, `negative`, `neutral`) are mapped to numerical IDs $\{0, 1, 2\}$ for model compatibility.

2. **Dialect-Aware Normalization**:

   - For **AraBERTv2**, we apply the official `ArabertPreprocessor` which handles Arabic-specific tokenization, diacritic removal, and light normalization while preserving dialectal markers.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 284–287
May 11, 2026.

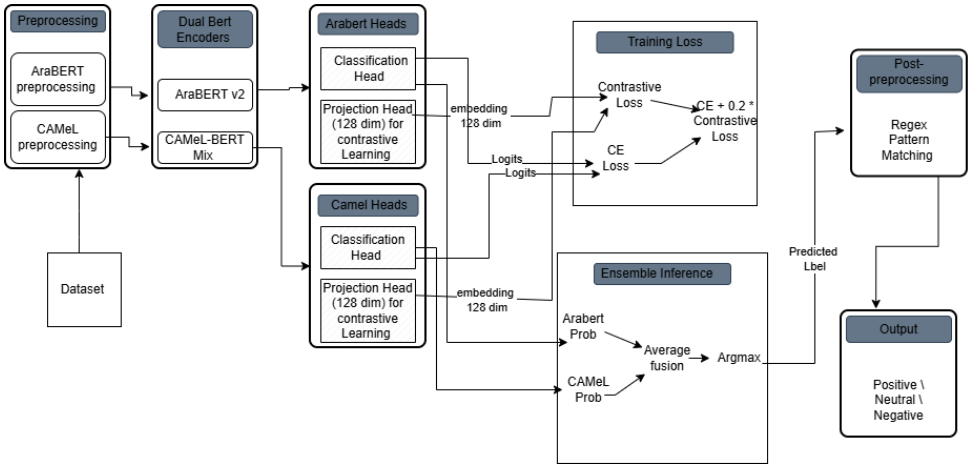

Figure 1: Workflow of the Proposed Ensemble Model for Arabic Sentiment Classification.

- For **CAMeLBERT**, we apply minimal preprocessing (whitespace trimming and null-value filtering) to retain the raw dialectal features the model was pre-trained on.

3. **Tokenization**: Each input is tokenized using the corresponding model's tokenizer with a maximum sequence length of 128 tokens, applying truncation and padding as needed.

We intentionally avoid aggressive preprocessing (e.g., stopword removal) to preserve sentiment-bearing dialectal expressions that transformer models can effectively learn from.

2.2. Base Models

We leverage two state-of-the-art pretrained Arabic transformers as our base learners, as summarized in Table 1.

Model	Identifier	Rationale
AraBERTv2	`aubmindlab/bert-base-arabertv2`	Trained on large Arabic corpora; strong generalization to MSA and light dialects.
CAMeLBERT-Mix	`CAMeL-Lab/bert-base-arabic-camelbert-mix`	Balanced MSA–dialect pretraining; effective for cross-dialect sentiment tasks.

Table 1: Pre-trained transformer models used in our system

Both models are extended with a custom classification head: a two-layer MLP with GELU activation, dropout ($p = 0.3$), and a final linear projection to 3 classes. Additionally, a projection head is attached for contrastive learning

2.3. Supervised Contrastive Learning Enhancement

To improve class separation in the embedding space, we integrate **Supervised Contrastive Loss** alongside the primary classification objective:

- A projection head maps the `[CLS]` token embedding to a 128-dimensional normalized space.
- During training, the total loss is computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{ConLoss}}$$

$$\mathcal{L}_{\text{ConLoss}} = \frac{1}{N} \sum_{i=1}^{N} \frac{-1}{|P(i)|} \log \left(\frac{\sum_{p \in P(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{j \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)} \right)$$

  where:

  - N is the number of valid anchor samples in the batch, i.e., samples having at least one positive pair.
  - $P(i) = \{p \in \{1, \ldots, B\} \setminus \{i\} \mid y_p = y_i\}$ is the set of indices of all samples in the batch sharing the same class label as anchor i.
  - $|P(i)|$ is the cardinality of the positive set for anchor i.
  - $\mathbf{z}_i \in R^{128}$ is the L2-normalized projected embedding of sample i, obtained through the projection head.
  - $\mathbf{z}_i \cdot \mathbf{z}_p$ denotes the dot product similarity between the embeddings of anchor i and positive sample p.
  - τ is the temperature scaling parameter controlling the sharpness of the distribution ($\tau = 0.07$ in experiments).
  - B is the total batch size, and the denominator sums over all samples $j \neq i$ in the batch, regardless of their label.

where $\lambda = 0.2$ and temperature $\tau = 0.07$.

- The contrastive term encourages embeddings of samples from the same sentiment class to cluster together while pushing apart embeddings from different classes, enhancing robustness to dialectal variation and ambiguous expressions.

2.4. Ensemble Strategy

Our ensemble combines predictions from both fine-tuned models using a **soft-voting** approach:

1. **Per-Model Inference**: Each model generates class probability distributions (via softmax over logits) for each input sample.
2. **Probability Averaging**: (Lakshminarayanan et al., 2017) Final probabilities are computed as the arithmetic mean of the two models' outputs:

$$P_{\text{final}} = \frac{P_{\text{AraBERT}} + P_{\text{CAMeLBERT}}}{2}$$

3. **Label Assignment**: The predicted label is obtained via $\arg\max$ over the averaged probabilities.

This strategy leverages the complementary strengths of both models: AraBERTv2's strong MSA foundation and CAMeLBERT's dialectal coverage, yielding more stable and generalizable predictions.

2.5. Rule-Based Post-Processing

To catch dialectal expressions that models often miss, we apply a lightweight rule-based post processing (Ray and Chakrabarti, 2022). Five pattern categories are checked via regex matching:

- **Neutral Triggers**: Dialectal phrases like (it's okay) override predictions to neutral for short texts.
- **Decline Indicators**: Temporal contrasts (e.g., "was excellent... became acceptable") downgrade positive to neutral.
- **Fake Review Detection**: Meta-commentary about review authenticity changes positive to negative.
- **Amenity Mentions**: References to facilities (clubs, bars, entertainment) at text start override to neutral.
- **Return Intent**: Expressions like "we will return" upgrade neutral to positive.

Rules are applied conditionally based on the model's initial prediction, ensuring minimal interference with high-confidence outputs.

3. Experimental Results

The following experiments evaluate our proposed approach on the AraSentEval 2026 Subtask 1 benchmark, including details on data, training configuration, and performance.

3.1. Dataset

The AraSentEval 2026 Subtask 1 dataset contains 1,731 Arabic hotel reviews across four dialects (Darija, Saudi, Jordanian, Egyptian), labeled as positive (609), negative (657), and neutral (465). Sentiment labels were mapped to numeric IDs, and each model used its own preprocessing pipeline and pretrained tokenizer with padding and truncation to a fixed maximum length. The data was split into training (85%) and validation (15%) sets with label stratification, without any external datasets or augmentation.

3.2. Training strategy

Each model is fine-tuned independently using AdamW optimizer with learning rate 2×10^{-5} for backbone parameters and 1×10^{-5} for classification/projection heads. a weight decay ($\lambda = 0.01$) and linear decay with 10% warmup steps. We fixed batch size to 16 (training), with gradient accumulation. In addition, we use Cross-Entropy with label smoothing ($\alpha = 0.1$) and class-weighted balancing to address label imbalance , and apply Gradient clipping ($\|\mathbf{g}\|_2 \leq 1.0$), dropout ($p = 0.3$), and early stopping (patience=5 epochs) based on validation macro-F1. Noting that all experiments were conducted on Google Colab with GPU acceleration and mixed-precision support.

3.3. Results

Our system was evaluated on both the validation set and the official AraSentEval 2026 Subtask 1 test set, with macro F1-score as the primary evaluation metric. During training, both individual models achieved strong validation performance:

- **AraBERTv2:** macro F1-score of 0.9355
- **CAMeLBERT-Mix:** macro F1-score of 0.9319

These results demonstrate the effectiveness of our training strategy. On the official competition test set, our ensemble system achieved:

- **Macro F1-score:** 0.8301
- **Accuracy:** 0.8301

The performance gap between validation and test sets is expected, given the diversity of Arabic dialects in the unseen test data. Overall, these results

validate our ensemble approach combining two Arabic transformers with supervised contrastive learning.

4. Conclusion

This paper presented our system for AraSentEval 2026 Subtask 1, focusing on robust sentiment analysis across diverse Arabic dialects. By combining two complementary pretrained Arabic transformers, AraBERTv2 and CAMeLBERT, within an ensemble framework enhanced by supervised contrastive learning, we aimed to address the linguistic variability and ambiguity inherent in dialectal Arabic sentiment classification. The integration of dialect-aware preprocessing, class-balanced training objectives, and lightweight rule-based post-processing further contributed to improving model robustness.

Our system achieved a macro F1-score of 0.83 on the official AraSentEval 2026 test set, confirming that contrastive representation learning and model ensembling are effective strategies for multi-dialect sentiment classification. While the gap between validation and test performance reflects the inherent difficulty of unseen dialectal expressions, it also motivates further research in this direction.

Looking ahead, we plan to explore data augmentation and cross-dialect alignment within the contrastive learning framework, as well as larger instruction-tuned Arabic language models. More adaptive post-processing and domain-specific pretraining are also promising directions to better capture the richness of Arabic user-generated content.

5. References

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. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)*, pages 9–15. European Language Resources Association. Model available at: https://huggingface.co/aubmindlab/bert-base-arabertv2.

Saad Ezzini, Shadi Abudalfa, Maram Alharbi, Salmane Chafik, Samaher Alghamdi, Reem Alotaibi, Hamzah Luqman, Mo El-Haj, and Paul Rayson. 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.

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 (WANLP 2021)*, pages 92–104. Association for Computational Linguistics. Model available at: https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc. Also available as arXiv:2004.11362.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Priyanka Ray and Amlan Chakrabarti. 2022. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*, 18(1/2):163–178.