TTLab at AraSentEval: SARF (صرف) Sentiment Analysis via Root-based Fusion for Multi-Dialectal Arabic

Ali Abusaleh, Bhuvanesh Verma, Prof. Dr. Alexander Mehler
Text Technoloy Lab (TTLab),
Goethe University Frankfurt
{a.abusaleh,verma,mehler}@em.uni-frankfurt.de

Abstract

Arabic sentiment analysis is challenged by morphological complexity and lexical variation across Arabic dialects, compounded by subjectivity in how speakers and writers express sentiment. In this paper, we present our submission for the AraSentEval 2026 Shared Task on Arabic Dialect Sentiment Analysis. We propose **SARF (صرف)** a multi-view architectural framework that integrates surface-level context with stemmed and rooted morphological perspectives using a shared MARBERTv2 encoder. Our system employs a hybrid BERT-CNN-BiLSTM-Attention architecture to capture both local sentiment n-grams and global sequential dependencies. Experimental results show that while individual morphological normalization strategies (stemming or rooting) may degrade performance, their joint integration via cross-morphological attention provides robust features across diverse dialects. Our final system achieved a competitive macro-F1-score of 0.9263, ranking 2nd out of 15 participating teams.

Keywords: NLP, Sentiment Analysis, Arabic analysis

1. Introduction

The advent of social media has fundamentally altered how individuals consume and produce information. Platforms such as X, Facebook, and Reddit are hubs for highly opinionated content that significantly influences public discourse. Similarly, online reviews on e-commerce websites provide personal insights and sentiments that drive consumer decision-making. In NLP, sentiment analysis is the task of polarity classification (where texts are typically labeled as positive, negative or neutral); sentiment analysis is treated as a core component of the broader area called opinion mining (Liu, 2012).

While sentiment analysis has been extensively researched across various languages, English remains the dominant focus due to its vast data availability (cf. Joshi et al. 2020). Despite being the fourth most used language on the web (Salloum et al., 2017), Arabic continues to trail behind English benchmarks in terms of state-of-the-art (SOTA) performance. A primary reason for this disparity is the prevalent use of informal Arabic dialects in online communication. While Modern Standard Arabic (MSA) is the formal register used in news, literature, and official documents, there are at least 22 distinct regional dialects used across the Arab world (Guellil and Azouaou, 2017). Because users interact on social media in their native dialects rather than MSA, Arabic text analysis presents distinct and substantial challenges for researchers.

To address these challenges, the Seventh Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) organized a shared task dedicated to Arabic Dialect Sentiment Analysis (Ezzini et al., 2026). This competition involves a multi-class classification challenge focused on identifying sentiment within texts written in various regional Arabic dialects. This task requires developing robust systems capable of navigating the substantial lexical and syntactic diversity found across the Arab world. To complete this task, we leveraged a state-of-the-art sentiment analysis framework based on the work of He and Abisado (2024). We extended this framework by introducing a novel architecture designed to incorporate morphological information specific to Arabic dialects (Procházka, 2021), thereby addressing their inherent linguistic complexity. Achieving a macro-F1-score of 0.9263, Our proposed system ranked 2nd out of 15 participating teams, demonstrating its effectiveness in handling dialectal variation.

2. Related work

While task-based sentiment analysis often struggles to generalize across domains or dialects (Shi and Agrawal, 2025), this challenge is particularly acute for Arabic, a highly context-sensitive language where meaning and sentiment can shift dramatically based on subsequent words or phrases. Bidirectional transformer models, such as BERT (Devlin et al., 2019), are naturally suited to this problem, as they capture context from both directions. This has been demonstrated in Arabic NLP through unitask and multitask models like HULMONA (ElJundi et al., 2019) and ARABERT (Antoun et al., 2020). The practical ef-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 262–268

fectiveness of such approaches is underscored by the 2021 Arabic Sentiment Competition at KAUST, where all top-ranking teams utilized variants of MARBERT (Abdul-Mageed et al., 2021; Alamro et al., 2021).

Pontiki et al. (2016) introduced the SemEval-2016 shared task on Aspect-Based Sentiment Analysis (ABSA). This task featured a specific track for Arabic hotel reviews, providing a dataset highly relevant to our task. Given the thematic overlap with our target domain, we leverage this corpus to enhance our model's training phase.

He and Abisado (2024) proposed a hybrid framework incorporating BERT, CNN, and BiLSTM layers to mitigate the effects of polysemy and improve feature representation. In their architecture, the CNN component is utilized to capture local, spatial features within the text, while the BiLSTM layer extracts global, sequential dependencies. We adopt and extend this hybrid methodology to effectively navigate the morphological and syntactic diversity across the Arabic dialectal landscape.

3. Task Description and Dataset

The primary objective of this shared task on Arabic Dialect Sentiment Analysis is to identify sentiment, categorized as positive, negative, or neutral, in texts written in various Arabic dialects. This problem is framed as a supervised multi-class classification challenge. Formally, given an input sequence S representing a sentence in a specific Arabic dialect, the model is required to learn a mapping function $f : S \rightarrow Y$. The label space Y is defined as the set of sentiment categories:

$$Y = \{\text{positive, negative, neutral}\} \quad (1)$$

The goal is to predict the most probable label $\hat{y} \in Y$ for each input S, formulated as:

$$\hat{y} = \arg\max_{y \in Y} P(y \mid S; \theta) \quad (2)$$

where θ represents the parameters of the classification model.

To facilitate the development of a robust classification model, the shared task utilizes the **Multi-Dialect-Sent (MDS-3)** dataset. This corpus consists of 3,000 sentences (1,731 in train set) annotated across three sentiment labels and covers four major regional dialects: Saudi (SA), Jordanian (JO), Moroccan (MA), and Egyptian (EG) (Alharbi et al., 2025c,a,b). The dataset was originally sourced from hotel reviews and subsequently translated and verified by native speakers of each respective dialect to ensure linguistic authenticity.

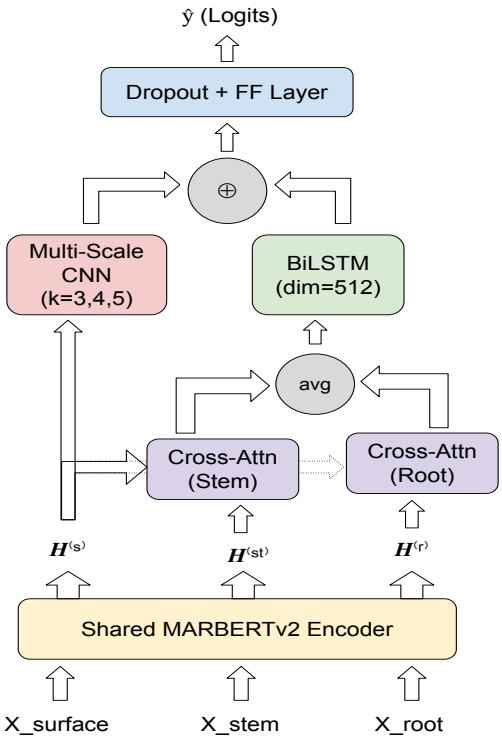

Figure 1: **SARF (صرف)** Architecture: A multi-view pipeline with shared MARBERTv2 encoder and CNN/BiLSTM dual-path.

4. Methodology

To develop a robust sentiment analysis system capable of addressing the dialectal nuances of Arabic, our approach integrates an established state-of-the-art (SOTA) framework with a novel architectural extension. This design specifically targets the inherent morphological and linguistic complexities present across diverse Arabic dialects. In the following subsections, we provide a detailed description of these primary components and our overall modeling strategy.

4.1. Cross morphological attention

We propose a multi-view architectural framework that captures linguistic nuances by representing a single input sentence through three parallel perspectives: the surface form, the stemmed form, and the rooted form. This approach ensures that the model maintains a balance between raw contextual semantics and underlying morphological structures.

4.1.1. Multi-View Encoding

Each linguistic view is first tokenized and subsequently mapped to a high-dimensional space using a shared pretrained transformer encoder $E(\cdot)$.

Category	Precision	Recall	F1-score
Dialect-Specific Ensemble System			
Darija	0.902	0.902	0.902
Egyptian	0.889	0.889	0.889
Jordanian	0.970	0.970	0.970
Saudi	0.967	0.967	0.967
Overall			
Ensemble System	0.9317	0.9317	0.9316
BERT-CNN-BiLSTM	0.9429	0.9421	**0.9425**
BERT-CNN-BiLSTM + Cross Morph. Attn	0.9398	0.9378	0.9388

Table 1: Performance of the Ensemble and BERT-CNN-BiLSTM-Self-Attention system. Dialect Specific results are using Ensemble system only.

Formally, let the input sequences be denoted as $X^{(s)}$, $X^{(st)}$, and $X^{(r)}$ for the surface, stemmed, and rooted views, respectively. The encoder generates contextualized representations as follows:

$$H^{(s)} = E(X^{(s)}),\ H^{(st)} = E(X^{(st)}),\ H^{(r)} = E(X^{(r)}) \tag{3}$$

where each representation $H^{(\cdot)} \in \mathbb{R}^{T \times d}$ corresponds to a sequence of length T with hidden dimension d.

4.1.2. Morphological Interaction and Fusion

To capture the dependencies between surface-level semantics and morphological abstractions, we employ a cross-attention mechanism. Here, the surface representation $H^{(s)}$ serves as the query, while the stemmed and rooted representations act as key-value pairs. This allows the model to align inflectional variations with their base forms:

$$\tilde{H}^{(st)} = \text{Attn}(H^{(s)}, H^{(st)}, H^{(st)}) \tag{4}$$

$$\tilde{H}^{(r)} = \text{Attn}(H^{(s)}, H^{(r)}, H^{(r)}) \tag{5}$$

The resulting features are then integrated into a unified representation $H^{(f)}$ through an element-wise averaging operation:

$$H^{(f)} = \frac{1}{3}\left(H^{(s)} + \tilde{H}^{(st)} + \tilde{H}^{(r)}\right) \tag{6}$$

4.2. BERT-CNN-BiLSTM

Building upon the framework of He and Abisado (2024), our model processes the surface level features $H^{(s)}$ and the morphologically enriched representation $H^{(f)}$ through parallel feature extraction pathways to capture both local and global linguistic patterns.

4.2.1. Feature Extraction Pathways

The local features, H_{local}, are extracted using a Convolutional Neural Network (CNN). For a filter W_c and bias b_c, the operation is defined as:

$$H_{local} = \text{ReLU}(W_c * H^{(s)} + b_c) \tag{7}$$

Simultaneously, global sequential dependencies are captured using a Bidirectional LSTM (BiLSTM). The representation H_{global} is the concatenation of the forward ($\overrightarrow{h}$) and backward ($\overleftarrow{h}$) hidden states:

$$H_{global} = [\overrightarrow{LSTM}(H^{(f)}) \oplus \overleftarrow{LSTM}(H^{(f)})] \tag{8}$$

4.2.2. Feature Fusion and Classification

The local and global features are subsequently fused via concatenation to form a comprehensive representation, $H_{total} = [H_{local} \oplus H_{global}]$. This composite vector is passed through a fully connected layer to compute the final sentiment distribution:

$$\hat{y} = \text{Softmax}(W_{fc} H_{total} + b_{fc}) \tag{9}$$

where W_{fc} and b_{fc} are the learnable weights and bias, respectively. The resulting output $\hat{y}$ represents the probability distribution over the label set $Y = \{\text{positive, negative, neutral}\}$, as illustrated in the full pipeline in Figure 1.

5. Experimental Setup

We utilize the SemEval-2016 Arabic reviews dataset, pre-training for 10 epochs and fine-tuning for 7 at a reduced learning rate (1.24×10^{-5}). Data preprocessing is performed using `PyArabic` (Zerrouki, 2010), `Tashaphyne` (Zerrouki, 2012b), and `Qalsadi` (Zerrouki, 2020, 2012a) to normalize the text, and extracting surface, stem, and root features. Our architecture employs a 128-dimensional LSTM and a 200-filter CNN (kernels: 3, 4, 5) with 0.3 dropout, 0.02 weight decay, and a batch size of 128 (8 for testing). The model is optimized using AdamW with an initial learning rate of 1.24×10^{-4}. Implementation details and the cleaning pipeline are available at GitHub.

6. Results and Discussion

6.1. Performance Overview

We evaluated **SARF** (صرف) against several competitive baselines, including an ensemble of SOTA Arabic transformers (MARBERTv2, AraBERT, and QARIB). As shown in Table 1, our hybrid **BERT-CNN-BiLSTM** approach achieved the highest performance on the AraSentEval DEVELOPMENT set with a **macro-F1-score of 0.9425**, outperforming the ensemble system (0.9316).

6.2. Synergy of Morphological Views

Our normalization results (Table 4) confirm that neither stemming nor rooting is a *silver bullet* in isolation. Rooting alone tends to remove sentiment-critical lexical nuances. However, the multi-view approach of **SARF** (صرف)–which aligns these views via cross-attention–successfully captures the stability of the root without losing the sentiment of the surface form. This confirms that the interaction between different morphological granularities is essential for handling the noise inherent in dialectal Arabic.

6.3. Final Shared Task Results

The final performance of our system on the "blind" test set provided by the ARASENTEVAL 2026 organizers is reported in Table 2.

Metric	Value
Macro Precision	0.9262
Macro Recall	0.9262
Macro-F1 Score	0.9262
Balanced Accuracy	0.8941

Table 2: Final results on the AraSentEval 2026 Shared Task test set.

7. Error Analysis

A consistent observation across all tested models and strategies was the significantly lower performance on the *Neutral* label. To investigate the root cause of this trend, we conducted a two-fold analysis: 1) a human re-annotation study of a sample training data and 2) an examination of model predictions. The methodology for the re-annotation is detailed in Appendix A.

7.1. Human Re-annotation and Label Noise

Our first analysis focused on assessing the quality of the original ground-truth labels through two distinct settings: **Relaxed Setting:** We analyzed all samples re-annotated by our human experts, regardless of the number of annotators per sample. Comparing these to the original training labels, we calculated a moderate Cohen's Kappa (κ) of 0.5360. Notably, of the 60 instances labeled as *Neutral* in the original dataset, our annotators re-classified 11 as *Negative* and 23 as *Positive*. **Strict Setting:** We restricted the analysis to 68 samples where consensus was reached by more than one annotator. In this setting, the inter-rater agreement with the original labels dropped significantly to a low $\kappa = 0.2958$. Among the 29 original *Neutral* samples in this subset, 15 were re-identified as *Positive* and 6 as *Negative*.

Sample	Original Label	Annotated Label
خدمة مزيانة، فندق ف موقع مزيان قريب من المطار المحلي و الدولي في أندهيري شرق مومباي، حنا عندنا *Good service, hotel in a good location near the local and international airport in Andheri East, Mumbai, we have...*	Neutral	Positive
السباحة كانت مزيانة و كان فيها حمام سباحة صغير و مشدود و فيه شلال فوسطها *The swimming was good; there was a small, well-maintained pool with a waterfall in the middle*	Neutral	Positive
الخدمة مزيانة، الأكل مزيان، الغرف مريحة، وخلال فترة وجودنا، شافنا فرح يقيم فالفندق، و كانو *The service was good, the food was good, the rooms were comfortable, and during our stay, we saw a wedding at the hotel*	Neutral	Positive

Figure 2: Re-annotation of *Hidden Positives*. These examples demonstrate how dialect-specific positive sentiment (such as the use of مزيانة for *good*) was originally overlooked and labeled as 'neutral', but accurately captured in our new annotations.

7.2. Model Prediction

Our error analysis suggests that the model internalized subjective, dialect-agnostic biases in the shared-task annotations. On instances where our re-annotations disputed the original labels, the model memorized the incorrect human bias 46.9% of the time, though its underlying pre-trained representations demonstrated some robustness by generalizing correctly in 40.6% of these cases. Crucially, this inherited noise disproportionately penalized specific regional dialects: while the model achieved moderate accuracy on the Jordanian (66.7%) and Saudi (64.7%) subsets, its performance degraded sharply to 44.4% on Darija. This degradation is directly linked to an inherited *safe bias* from the original annotators, where the model most frequently misclassified both positive and negative instances as Neutral (accounting for 86% and 100% of their respective errors).

Figure 2 shows the original and new annotations for a few samples. This discrepancy indicates that the model's suboptimal performance on neutral text is likely due to false signals and label

noise within the training data, rather than architectural limitations.

8. Conclusion

In this paper, we described our submission to the AraSentEval 2026 Shared Task on Arabic Dialect Sentiment Analysis. Our system, **SARF (صرف)**, addresses the morphological complexity of 4 regional dialects by integrating surface-level context with stemmed and rooted views via a cross-attention mechanism. By building a hybrid BERT-CNN-BiLSTM-Attention framework on top of shared MARBERTv2 encoder, we successfully captured both local sentiment n-grams and global sequential dependencies. Our results demonstrate that while individual morphological normalization (rooting) can be reductive, a multi-view approach provides a robust representation that is invariant to dialectal noise. Furthermore, our error analysis revealed significant label noise in the *Neutral* class, suggesting that the current performance ceiling in Arabic Dialect Sentiment Analysis may be limited more by annotation consistency than by model architecture. Our system ranked 2nd, indicating that morphologically aware hybrid models can be effective for low-resource, high-variance dialectal tasks.

9. Limitations & Ethical Considerations

Limitations Our system faces three primary limitations. First, the **SARF (صرف)** architecture requires three parallel encoding passes (surface, stem, and root), which triples the computational overhead during inference compared to standard Transformer models. This may limit scalability for massive real-time streams. Second, our reliance on the `Qalsadi` and `Tashaphyne` tools for rooting means that errors in the rule-based morphological analyzer can propagate through the cross-attention layers. Finally, while we investigated four major dialects, the model's performance on more localized or "deep" dialectal expressions (especially in the Maghrebi cluster) remains lower than its performance on Levantine and Gulf dialects, likely due to the linguistic distance from the MARBERTv2 pre-training corpus.

Ethical Considerations Our error analysis specifically identified a "safe bias" in the original annotations that disproportionately mislabeled Darija (Moroccan) expressions. We emphasize that the systemic misclassification of specific regional dialects can lead to linguistic exclusion in AI-driven services. To mitigate this, we advocate for the inclusion of native speakers from across the Mashreq (المشرق) and Maghreb (المغرب) in future annotation efforts. Our code and fine-tuned weights are released for academic use to promote transparency in dialectal NLP research.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Hind Alamro, Manal Alshehri, Basma Alharbi, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2021. Overview of the arabic sentiment analysis 2021 competition at kaust.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hULMonA: The universal language model in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77, Florence, Italy. Association for Computational Linguistics.

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.

Imène Guellil and Faiçal Azouaou. 2017. Asda: Analyseur syntaxique du dialecte alg {\'e} rien dans un but d'analyse s {\'e} mantique. *ArXiv preprint*, abs/1707.08998.

Aixiang He and Mideth Abisado. 2024. Text sentiment analysis of douban film short comments based on bert-cnn-bilstm-att model. *Ieee Access*, 12:45229–45237.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment analysis and opinion mining*, volume 5 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Stephan Procházka. 2021. *Arabic Dialectology*, Cambridge Handbooks in Language and Linguistics, page 214–243. Cambridge University Press.

Said A Salloum, Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. 2017. Using text mining techniques for extracting information from research articles. In *Intelligent natural language processing: Trends and Applications*, pages 373–397. Springer.

Zhiqiang Shi and Ruchit Agrawal. 2025. A comprehensive survey of contemporary Arabic sentiment analysis: Methods, challenges, and future directions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3760–3772, Albuquerque, New Mexico. Association for Computational Linguistics.

Taha Zerrouki. 2010. pyarabic, an arabic language library for python.

Taha Zerrouki. 2012a. qalsadi, arabic mophological analyzer library for python.

Taha Zerrouki. 2012b. Tashaphyne, arabic light stemmer.

Taha Zerrouki. 2020. Towards an open platform for arabic language processing.

A. Annotation Methodology

To validate the quality of the original labels, we randomly sampled 30 instances per dialect from the training set. The re-annotation task was performed by native speakers of the respective dialects, all of whom were university students or graduates. The annotator pool ranged in age from 22 to 50 years, with a mean age of 29.

Dialect	Pos	Neg	Neu
Moroccan (Darija)	17	5	5
Saudi	26	21	18
Jordanian	45	45	28

Table 3: Distribution of sentiment labels across the re-annotated dialects.

A.1. Annotator Distribution and Data Collection

The distribution of annotators was as follows: 11 for Jordanian, 5 for Saudi, 5 for Moroccan (Darija), and 2 for Egyptian. Due to a lack of completed responses from the Egyptian cohort, this dialect was excluded from the subsequent analysis. Following the collection phase, the final re-annotated dataset comprised 119 instances for Jordanian, 68 for Saudi, and 30 for Moroccan.

A.2. Consensus and Analysis Settings

To determine the final labels for our study, we applied a majority voting scheme to each instance. We then categorized the data into two analytical settings:

- **Strict Analysis:** Only instances with two or more annotations were considered to ensure higher label confidence.
- **Relaxed Analysis:** All instances with at least one expert annotation were included.

The resulting label distributions and a comparison with the original task labels are summarized in Table 3.

B. Language-specific Normalization

As shows in Table 4, across both models, applying stemming or rooting independently does not yield consistent improvements over the baseline without normalization. In particular, rooting alone leads to a noticeable degradation in performance, indicating that collapsing words to their morphological roots removes sentiment relevant lexical distinctions. Similarly, stemming alone shows marginal differences relative to the baseline, suggesting limited added value in the presence of strong pretrained representations. In contrast, combining stemming and rooting consistently improves performance for both MARBERT and MARBERTv2. These findings suggest that multi-granularity morphological normalization allows the model to benefit from **dialect-level grouping via roots** while preserving surface-level and contextual information necessary for sentiment discrimination.

Setup	Macro-F1	Accuracy	Macro Precision	Macro Recall
		MARBERT		
No stem, no root	0.8658	0.8719	0.8692	0.8656
Stem only	0.8627	0.8686	0.8645	0.8626
Root only	0.8522	0.8587	0.8544	0.8527
Stem + root	**0.8762**	**0.8818**	**0.8766**	**0.8770**
		MARBERTv2		
No stem, no root	0.8838	0.8884	0.8840	0.8850
Stem only	0.8822	0.8876	0.8849	0.8819
Root only	0.8813	0.8868	0.8822	0.8815
Stem + root	**0.8951**	**0.9000**	**0.8959**	**0.8951**

Table 4: Cross-validated performance on AraSentEval **Validation-dataset** under different normalization strategies.