Comparative Study of Machine Learning and Transformer-Based Approaches for Arabic Politeness Detection at AdabEval 2026

Mariem Ben Arbia[1], **Ghada Ben Amor**[1], **Houssem Safi**[2], **Omar Trigui**[1,2]
[1]The Higher Institute of Management of Sousse (ISGS), University of Sousse, Tunisia
[2]ANLP Research Group, MIRACL Laboratory, Sfax, Tunisia

Abstract

This paper describes our system submitted to the OSACT7 AdabEval shared task on Arabic politeness detection (Task A). The task requires classifying Arabic texts into three categories: *Polite*, *Impolite*, and *Neutral*. We systematically explore multiple approaches, progressing from classical machine learning baselines using pre-trained embeddings to fine-tuned transformer models. Our best system leverages MARBERT, a transformer model pre-trained on one billion Arabic tweets, fine-tuned with Focal Loss to handle the significant class imbalance present in the dataset (∼75% Neutral). We additionally experiment with hybrid approaches combining fine-tuned embeddings with gradient-boosted classifiers and ensemble methods. Our best single model achieves a macro F1 score of 0.84 and an accuracy of 0.90 on the validation set, substantially outperforming classical ML baselines (F1 = 0.42).

Keywords: Arabic NLP, politeness detection, MARBERT, Focal Loss, text classification, shared task

1. Introduction

Politeness detection in natural language is an important task for understanding social dynamics in online communication (Danescu-Niculescu-Mizil et al., 2013). While significant progress has been made for English, Arabic politeness detection remains underexplored despite the language's rich pragmatic features and wide dialectal variation (Brown and Levinson, 1987).

The OSACT7 AdabEval shared task (Alqifari et al., 2026) addresses this gap by proposing a three-way classification task: given an Arabic text (tweet, product review, or YouTube comment), the system must predict whether the text is *Polite*, *Impolite*, or *Neutral*. This task presents several challenges:

- **Class imbalance**: The dataset is heavily skewed toward the *Neutral* class (∼75%), with *Polite* representing ∼17% and *Impolite* only ∼8.5%.
- **Dialectal diversity**: The data includes Modern Standard Arabic (MSA) as well as multiple Arabic dialects (Egyptian, Saudi, Gulf, etc.).
- **Implicit politeness**: Politeness cues in Arabic can be subtle and context-dependent, involving honorifics, religious expressions, and indirect speech acts.

In this paper, we describe our systematic exploration of multiple approaches, from feature-based classical machine learning to fine-tuned transformers. We show that the choice of pre-trained model is critical: MARBERT (Abdul-Mageed et al., 2021), trained on Arabic tweets, dramatically outperforms AraBERT (Antoun et al., 2020), which was primarily trained on formal Arabic text. We further demonstrate that Focal Loss (Lin et al., 2017) is more effective than standard cross-entropy for handling the class imbalance inherent in this task.

2. Related Work

Computational approaches to politeness detection were pioneered by Danescu-Niculescu-Mizil et al. (2013), who developed a framework for English based on linguistic features derived from politeness theory (Brown and Levinson, 1987). For Arabic, the field is comparatively nascent, with most work focusing on sentiment analysis rather than politeness per se.

Pre-trained language models have become the dominant approach for Arabic NLP tasks. AraBERT (Antoun et al., 2020) was trained on news articles and Wikipedia, making it well-suited for MSA but less effective on dialectal text. MARBERT (Abdul-Mageed et al., 2021) addresses this limitation by pre-training on one billion Arabic tweets, capturing dialectal patterns and informal language use. The effectiveness of domain-matched pre-training for downstream tasks has been widely documented in the literature.

Focal Loss (Lin et al., 2017), originally proposed for object detection, has been successfully applied to text classification tasks with imbalanced data. By down-weighting well-classified examples, it allows the model to focus on hard, ambiguous cases — a particularly relevant property for politeness detection where the boundaries between classes can be subtle.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 179–184
May 11, 2026.

3. Data

The AdabEval dataset for Task A (Al-Khalifa et al., 2026) consists of Arabic texts from three sources: Twitter, company reviews, and Shein product reviews. Each text is labeled as *Polite*, *Impolite*, or *Neutral*.

Split	Total	Neut.	Pol.	Imp.
Train	4,895	74.9%	16.6%	8.5%
Val	693	74.9%	16.6%	8.5%
Test	1,406	—	—	—

Table 1: Dataset statistics for Task A. The test set labels were not released during the competition.

As shown in Table 1 and Figure 1, the dataset exhibits a strong class imbalance, with the *Neutral* class representing approximately 75% of the data. This imbalance motivates our use of specialized loss functions and resampling techniques.

4. Methodology

We explore three families of approaches, each building upon insights from the previous one. Figure 2 illustrates the architecture of our best-performing system.

4.1. Text Preprocessing

All approaches share a common preprocessing pipeline:

1. Removal of URLs, user mentions (@), and hashtags (#).
2. Arabic character normalization: all variant forms of *alef* are unified to bare *alef*, *ta marbuta* is mapped to *ha*, and *alef maqsura* to *ya*.
3. Whitespace normalization.

For AraBERT-based experiments, we additionally apply the AraBERT preprocessor, which performs segmentation and further normalization steps specific to that model.

4.2. Approach 1: Feature Extraction with Classical ML

Our first approach uses pre-trained AraBERT v2 as a fixed feature extractor. We extract [CLS] token embeddings (768 dimensions) from the model without any fine-tuning. These embeddings are then used to train several classical ML classifiers:

- Logistic Regression
- Random Forest
- XGBoost (Chen and Guestrin, 2016)
- Support Vector Machine (SVM)

This approach serves as our baseline, establishing the performance achievable without task-specific model adaptation.

4.3. Approach 2: Fine-Tuning with Cross-Entropy

Building on the baseline results, we fine-tune AraBERT v2 end-to-end with a classification head and standard weighted cross-entropy loss. We experiment with multiple learning rates and training configurations.

4.4. Approach 3: MARBERT with Focal Loss

Our best approach makes two key modifications:

Model selection. We replace AraBERT with MARBERT (Abdul-Mageed et al., 2021), which was pre-trained on one billion Arabic tweets. Given that our dataset includes tweets and informal reviews, this model provides a better domain match.

Loss function. We replace weighted cross-entropy with Focal Loss (Lin et al., 2017), defined as:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the predicted probability for the correct class, $\gamma = 2.0$ is the focusing parameter that down-weights easy examples (this value follows the recommendation of the original paper (Lin et al., 2017)), and α_t represents per-class weights computed inversely proportional to class frequency.

Training details. We fine-tune MARBERT with the following hyperparameters. The learning rate and weight decay follow standard recommendations for BERT-style fine-tuning; the batch size and sequence length were chosen to fit GPU memory constraints; early stopping patience was set to 4 based on preliminary development set experiments:

- Learning rate: 1×10^{-5} with linear warmup (10% of total steps)
- Optimizer: AdamW (Loshchilov and Hutter, 2019) with weight decay 0.01
- Batch size: 16
- Maximum sequence length: 128 tokens
- Gradient clipping: max norm 1.0
- Early stopping: patience of 4 epochs
- Maximum epochs: 15

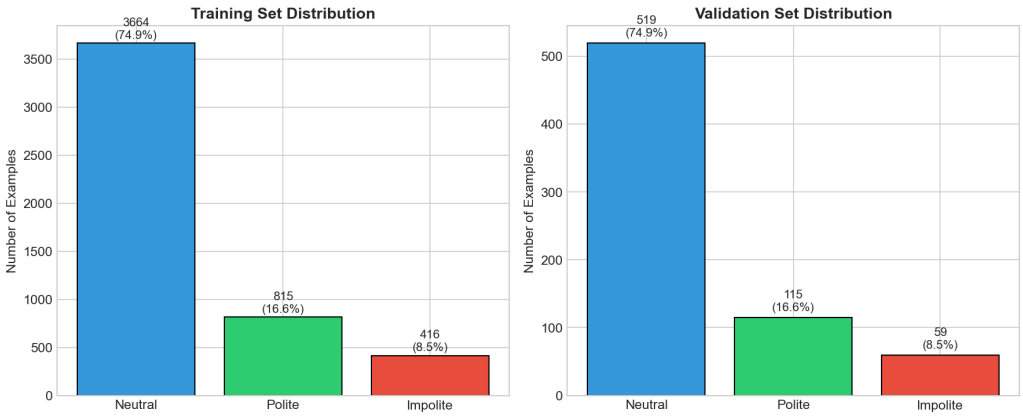

Figure 1: Class distribution in the training and validation sets. The *Neutral* class dominates both splits, representing approximately 75% of the data.

4.5. Approach 4: Hybrid (Fine-Tuned Embeddings + XGBoost)

We also explore a hybrid approach: extracting [CLS] embeddings from the *fine-tuned* MARBERT model and using them as features for an XGBoost classifier (Chen and Guestrin, 2016). To address class imbalance at the feature level, we apply SMOTE (Chawla et al., 2002) to the training embeddings before training XGBoost.

4.6. Approach 5: Ensemble

Finally, we experiment with soft-voting ensembles combining predictions from multiple approaches. We test two ensemble configurations:

- **All**: Combining all four individual models.
- **FT + Hybrid**: Combining the best fine-tuned model with the best hybrid model.

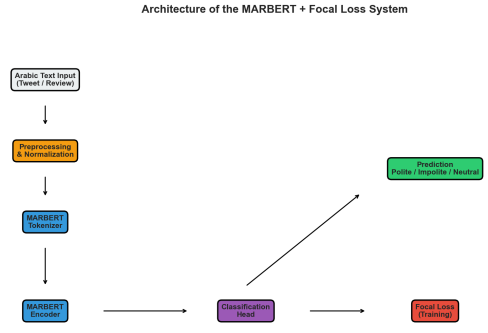

Figure 2: Architecture of our best system: MARBERT fine-tuned with Focal Loss for three-way politeness classification.

5. Results

5.1. Comparison of Approaches

Table 2 presents the results of all approaches on the validation set. The comparison is also visualized in Figure 3.

Approach	Acc	P	R	F1
Classical ML				
AraBERT+XGB	.69	—	—	.42
Fine-tuning				
AraBERT v2	.30	.40	.38	.27
MARBERT+FL	**.90**	**.82**	**.88**	**.84**
Hybrid				
MARB.→XGB	.68	.32	.33	.32
AraB.→XGB	.49	.33	.33	.32
Ensembles				
All models	.83	.81	.57	.64
FT+Hybrid	.86	.76	.74	.75

Table 2: Results on the validation set. P = Macro Precision, R = Macro Recall, F1 = Macro F1. FL = Focal Loss, XGB = XGBoost+SMOTE. For Classical ML, all four classifiers (LR, RF, SVM, XGBoost) were evaluated; XGBoost achieved the best macro F1 (0.42) and is reported here as the representative baseline.

The results reveal several important findings:

MARBERT vs. AraBERT. MARBERT fine-tuning achieves F1 = 0.84, while AraBERT v2 fine-tuning yields only F1 = 0.27. This dramatic difference highlights the importance of domain-matched pre-training: MARBERT, trained on Arabic tweets, is far better suited for informal text classification than AraBERT, trained primarily on news and Wikipedia.

Fine-tuning vs. Feature Extraction. The classical ML baseline using frozen AraBERT embed-

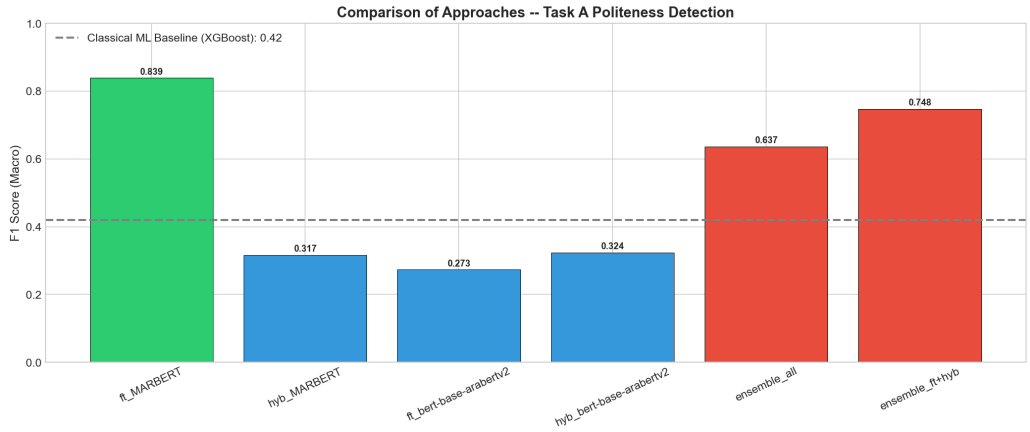

Figure 3: Macro F1 scores across all approaches on the validation set. Bars are colour-coded: green = best fine-tuned model (MARBERT + Focal Loss); red = ensemble variants; blue = all other approaches (fine-tuning baselines, hybrid models). The dashed grey line marks the classical ML baseline (F1 = 0.42).

dings (F1 = 0.42) actually outperforms AraBERT fine-tuning (F1 = 0.27), suggesting that fine-tuning AraBERT on this small, domain-mismatched dataset leads to catastrophic forgetting or poor convergence.

Focal Loss. MARBERT with Focal Loss achieves balanced performance across all three classes, with the focusing mechanism ($\gamma = 2.0$) effectively preventing the model from being dominated by the majority *Neutral* class. While we did not conduct a direct comparison of AraBERT with Focal Loss (which would isolate the effect of the loss function from the model choice), the large gap between AraBERT fine-tuning (F1 = 0.27) and MARBERT fine-tuning (F1 = 0.84) is primarily attributable to domain mismatch rather than the loss function. The choice of MARBERT over AraBERT is motivated by its pre-training on Arabic social media data, and the choice over other Arabic BERT-based models by its demonstrated state-of-the-art performance on dialectal Arabic tasks (Abdul-Mageed et al., 2021).

Hybrid approaches. Extracting embeddings from fine-tuned models and training XGBoost + SMOTE (F1 = 0.32) performs worse than direct fine-tuning, likely because the dimensionality reduction loses task-relevant information captured in the attention layers.

Ensembles. Ensemble methods do not improve upon the best single model, as the weaker models introduce noise that degrades overall performance.

5.2. Per-Class Analysis

Table 3 shows the detailed per-class results for our best model (MARBERT + Focal Loss).

The model achieves strong performance across all three classes. The *Neutral* class benefits from its larger representation (F1 = 0.94), and the minority classes achieve F1 scores of 0.85 for *Polite* and 0.75 for *Impolite*. The *Impolite* class is the most challenging, which is expected given both the diversity of expressions of impoliteness across Arabic dialects and the fact that it is the most underrepresented class (∼8.5% of training data).

Class	Prec.	Rec.	F1	Sup.
Impolite	.70	.81	.75	59
Neutral	.96	.91	.94	519
Polite	.79	.90	.85	115
Macro Avg	.82	.88	.84	693

Table 3: Per-class results of the best model (MARBERT + Focal Loss) on the validation set.

Figure 4 shows the confusion matrix for our best model. The main error patterns are: 20 *Neutral* examples misclassified as *Impolite*, 25 *Neutral* examples misclassified as *Polite*, and 9 *Impolite* examples misclassified as *Neutral*, reflecting the inherent ambiguity at the boundary between neutral and impolite speech.

5.3. Test Set Predictions

Figure 5 compares the distribution of our MARBERT model's predicted labels on the test set against the training set label distribution. Since the ground-truth labels for the test set were not released during the competition, this comparison does not measure prediction accuracy but rather checks whether the model's output distribution is consistent with the expected class proportions from training.

Our MARBERT model produces a prediction distribution on the test set that closely mirrors the training data proportions, suggesting well-calibrated predictions and the absence of majority-class collapse.

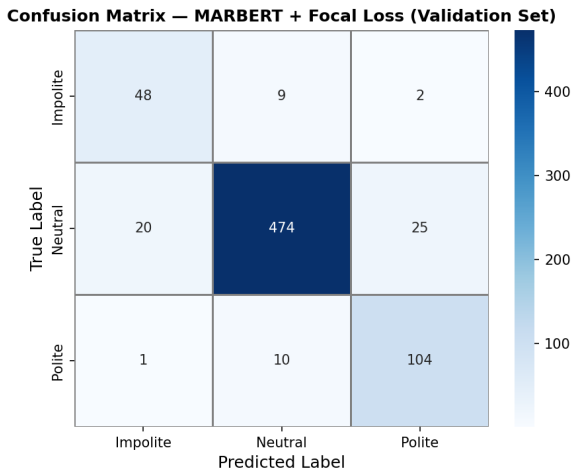

Figure 4: Confusion matrix of the best model (MARBERT + Focal Loss) on the validation set (693 samples).

This is in contrast to weaker baselines which tend to over-predict the majority *Neutral* class.

6. Discussion

Our experiments highlight several key lessons for Arabic text classification in the context of social media:

Pre-training domain matters more than model architecture. The most impactful decision in our pipeline was switching from AraBERT (news/Wikipedia) to MARBERT (tweets). Both models share a similar BERT architecture, but MARBERT's pre-training on 1 billion Arabic tweets provides it with knowledge of dialectal expressions, informal spelling, and social media conventions that are critical for politeness detection in this domain.

Focal Loss outperforms resampling. While SMOTE (Chawla et al., 2002) is a popular approach for handling class imbalance, we find that Focal Loss applied directly during fine-tuning is more effective. This is likely because Focal Loss operates at the loss level, allowing the model to learn nuanced decision boundaries, whereas SMOTE operates at the data level and can introduce noise through synthetic interpolation in high-dimensional embedding spaces.

Simple models can outperform complex pipelines. Our best result comes from a straightforward fine-tuning approach (MARBERT + classification head + Focal Loss), without any additional architectural modifications or multi-stage pipelines. Hybrid approaches and ensembles did not improve performance, reinforcing the principle that well-matched pre-training combined with appropriate loss functions can be sufficient.

7. Conclusion

We presented our approach to the OSACT7 AdabEval Task A on Arabic politeness detection. Through systematic experimentation, we demonstrated that MARBERT fine-tuned with Focal Loss achieves strong results (F1 = 0.84), representing a 100% improvement over classical ML baselines. Our findings emphasize the critical importance of domain-matched pre-training for Arabic NLP tasks involving informal text. The key factors behind our system's success are: (1) using a model pre-trained on data similar to the target domain, (2) employing Focal Loss to handle class imbalance at the loss level, and (3) careful regularization through early stopping and gradient clipping.

For future work, we plan to explore data augmentation techniques for the minority classes, investigate multi-task learning with related Arabic NLP tasks, and experiment with larger language models. We also plan to investigate the role of dialectal features in politeness classification through per-source analysis.

8. Limitations

Our work has several limitations. First, our validation results may not fully generalize to the test set, as the official test labels were not available during development. Second, the AraBERT fine-tuning results (F1 = 0.27) are surprisingly low — lower even than the frozen AraBERT baseline (F1 = 0.42). We attribute this anomaly to a combination of domain mismatch (AraBERT was pre-trained on news and Wikipedia, while our data contains informal social media text) and potential optimization instability: fine-tuning a mismatched model on a small, imbalanced dataset can lead to catastrophic forgetting or convergence to a degenerate solution. However, we cannot exclude the possibility of a hyperparameter or implementation issue, and a more thorough ablation (e.g., with learning rate sweep and class-weighted loss for AraBERT) would be needed to fully rule this out. Third, we did not perform extensive hyperparameter search due to computational constraints, and our MARBERT results could potentially be improved further with systematic tuning.

9. Bibliographical 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*

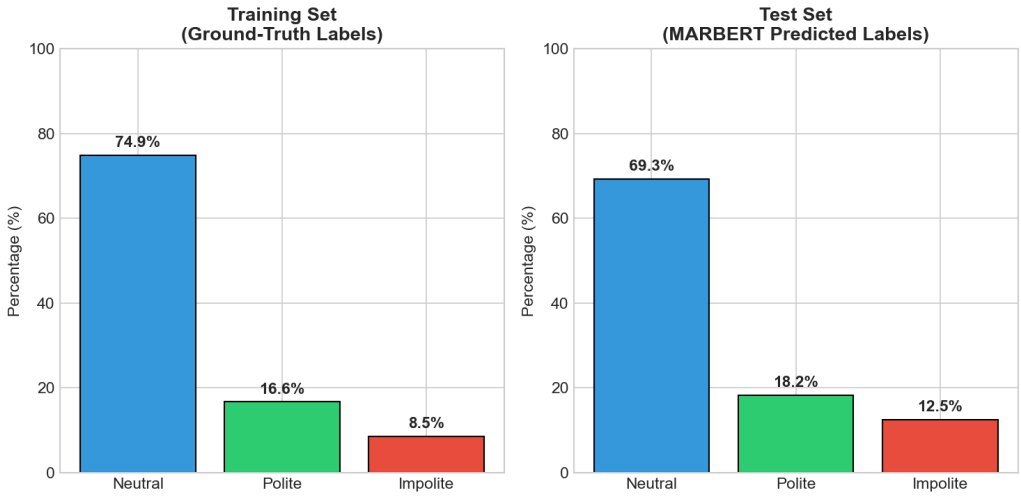

Figure 5: Comparison of class label distributions: training set (left, ground-truth labels) vs. MARBERT predictions on the test set (right, predicted labels). The similarity between the two distributions suggests that the model is well-calibrated and does not collapse to predicting the majority class. Note that ground-truth test labels were unavailable; this figure reflects the model's predictions only.

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7088–7105. Association for Computational Linguistics.

Hend Al-Khalifa, Nadia Ghezaiel, Maria Bounnit, Hend Hamed Alhazmi, Noof Abdullah Alfear, Reem Fahad Alqifari, Ameera Masoud Almasoud, and Sharefah Al-Ghamdi. 2026. ADAB: Arabic dataset for automated politeness benchmarking — a large-scale resource for computational sociopragmatics. In *Proceedings of The 2026 International Conference on Language Resources and Evaluation (LREC2026)*, Palma, Mallorca (Spain).

Reem Alqifari, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Maria Bounnit, Hend AlHazmi, Ameera Almasoud, Sharefah AlGhamdi, and Noof Alfear. 2026. The AdabEval 2026 shared task on Arabic politeness detection. In *Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks (OSACT 2026) co-located with the 2026 International Conference on Language Resources and Evaluation (LREC2026)*, Palma, Mallorca (Spain).

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*, pages 9–15.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4 of *Studies in Interactional Sociolinguistics*. Cambridge University Press.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cristian Danescu-Niculescu-Mizil, Moritz Suber, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.