Multi-Label Classification of Arabic Politeness Criteria in Social Media

Rand Abdullah Alturki
King Saud University
Riyadh, Saudi Arabia
rand.a.t.alturki@gmail.com

Abstract

We address the problem of multi-label classification of politeness and impoliteness criteria in Arabic social media posts, as defined in Subtask B of an Arabic politeness shared task. The goal is to assign up to four labels from nine pragmatic categories, including Insult, Criticism, Respect, Prayers, and Hospitality, to each post. We first construct consistent multi-label annotations by mapping heterogeneous criterion strings into the official label set and analyzing their skewed distribution. To mitigate severe class imbalance, especially for rare categories such as Hospitality and Racism/Discrimination, we apply targeted oversampling of minority instances. Our modelling pipeline combines a TF–IDF + Logistic Regression baseline with two transformer-based encoders, MARBERT and AraBERT-twitter, trained for multi-label classification with Focal Loss. We then aggregate model outputs through a weighted ensemble and optimize per-class decision thresholds on a held-out validation set to improve macro-averaged F1. Experiments on the shared-task train/validation split show that the ensemble substantially outperforms the TF–IDF baseline and individual transformers, particularly on underrepresented categories, while maintaining competitive performance on frequent labels.

Keywords: Arabic, politeness, impoliteness, multi-label classification, social media, transformers

1. Introduction

Politeness and impoliteness are crucial aspects of interpersonal communication, and modeling them in Arabic social networks is important for applications such as content moderation, conversational agents, and sociolinguistic analysis. Subtask B of the Arabic politeness shared task (Alqifari et al., 2026) focuses on assigning up to four pragmatic criteria to each tweet, drawn from a fixed inventory of nine categories such as Insult, Criticism, Respect, Prayers, and Hospitality. Unlike standard single-label sentiment classification, this setup is inherently multi-label and exhibits strong class imbalance, with some categories occurring only a handful of times in the training data. These characteristics make the task challenging: systems must capture fine-grained pragmatic distinctions while remaining robust to sparse supervision for rare labels.

In this paper, we present a complete multi-label classification pipeline for Subtask B. We unify noisy annotation strings into the official nine categories, analyze category distributions before and after balancing, and train both traditional and neural models. Our approach combines a TF–IDF + Logistic Regression baseline with two Arabic transformer encoders, MARBERT (Ezzeldin et al., 2025) and AraBERT-twitter (Hithnawi et al., 2025), using Focal Loss (202, 2024) to better handle label imbalance. We then build a three-way ensemble and tune per-class decision thresholds to optimize macro-averaged F1 on a held-out validation set. The proposed system is simple to reproduce, yet delivers strong performance on both frequent and rare categories.

2. Related Work

Politeness and impoliteness in computational linguistics have been studied primarily for English, often using single-label classification of requests, replies, or forum posts. Work on Arabic pragmatics has mainly focused on sentiment, toxicity, and hate speech (Alqifari et al., 2026), with relatively less attention to fine-grained politeness criteria in social media. Multi-label text classification is a well-established problem, and prior work shows that combining sparse lexical models such as TF–IDF with transformer-based encoders can yield robust performance, especially under label imbalance (Liang et al., 2025).

For Arabic social media, MARBERT (Ezzeldin et al., 2025) and AraBERT-family models (Hithnawi et al., 2025) have become strong backbones for a range of tasks, including sentiment analysis, offensive language detection, and dialect identification (Alqifari et al., 2026). Our work extends these lines by targeting multi-label politeness criteria and carefully addressing label skew through oversampling and Focal Loss (202, 2024), combined with a simple yet effective ensemble.

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 185–190

3. Task and Data

3.1. Task Definition

Subtask B requires assigning up to four pragmatic criteria to each Arabic tweet. The official inventory consists of nine categories: Criticism, Insult, Respect, Prayers, Greetings, Hospitality, Gratitude, Admiration / Love, and Racism / Discrimination. Each text may be associated with one to four criteria, annotated in separate columns ("criteria 1"–"criteria 4") in the provided CSV files. The evaluation metric is macro-averaged F1 over the nine categories.

3.2. Dataset

The organizers provided us with three data splits: training, validation, and test, all with the same column structure. We focus on the train and validation sets for model development and use the test set only for final submission. The text is stored in an unnamed first column, followed by a label for politeness, and then up to four criteria columns with their associated key-word fields. For Subtask B we ignore the politeness label and key-word columns, using only the tweet text and criteria annotations.

Table 1 reports the per-category counts in the original training set, based on the unified nine-category mapping described below. The distribution is highly skewed: Respect and Insult are common, while Hospitality and Racism / Discrimination are extremely rare.

Category	Train	Val
Criticism	207	30
Insult	514	74
Respect	787	110
Prayers	415	59
Greetings	93	16
Hospitality	3	1
Gratitude	271	29
Admiration / Love	256	39
Racism / Discrimination	10	1

Table 1: Category counts in original train/validation sets (before balancing).

3.3. Label Mapping

The raw criteria columns contain a variety of strings, including bilingual labels (e.g., "Admiration & Love - الحب و الإعجاب" *al-i'jāb wa al-ḥub*) and fine-grained distinctions (e.g., "Verbal violence - عنف لفظي" *'unf lafẓī*, "Threat - تهديد" *tahdīd*). We construct a mapping dictionary that collapses all observed variants into the nine official categories as shown in Table 2. For example, we map *Disparagement*, *Verbal violence* - عنف لفظي *('unf lafẓī)*, *Threat* - تهديد *(tahdīd)*, and *Sarcasm* - تهكم *(tahakkum)* into *Insult*, while multiple "Asking for permission" variants are mapped into *Respect*. We then parse each row by scanning the four criteria columns, applying the mapping, and collecting the resulting labels into a list.

Raw Annotation String	Mapped Category
Insult	
Disparagement	
Verbal violence (عنف لفظي, *'unf lafẓī*)	
Threat (تهديد, *tahdīd*)	
Sarcasm (تهكم, *tahakkum*)	Insult
Criticism (انتقاد, *intiqād*)	
Accusation	Criticism
Respect	
Asking for permission (طلب الإذن, *ṭalab al-idhn*)	
Asking for permission & affection (طلب الإذن و التودد, *ṭalab al-idhn wa al-tawadud*)	
Excuse (اعتذار, *i'tidhār*)	Respect
Greetings (التحية, *al-taḥiyya*)	Greetings
Hospitality & generosity (الكرم و الضيافة, *al-ḍiyāfa wa al-karam*)	Hospitality
Prayers	Prayers
Gratitude (الامتنان, *al-imtinān*)	
Thanks & gratitude (الامتنان و الشكر, *al-shukr wa al-imtinān*)	
Congratulations (تهنئة, *tahni'a*)	
Felicitation (التهنئة, *al-tahni'a*)	Gratitude
Admiration (الإعجاب, *al-i'jāb*)	
Love (الحب, *al-ḥub*)	
Appreciation & Love (الحب و الإعجاب, *al-i'jāb wa al-ḥub*)	Admiration / Love
Racism & discrimination (تمييز و عنصرية, *'unṣuriyya wa tamyīz*)	
Discrimination & sectarian (عنصرية، تمييز و تحزب, *'unṣuriyya, taḥazzub wa tamyīz*)	Racism / Discrimination

Table 2: Mapping of raw annotation strings to the nine official categories. Multiple surface variants (including bilingual labels) are collapsed into a single canonical label.

Text with no valid criteria are represented by an empty label list. We use a MultiLabelBinarizer with a fixed class order to convert the label lists into 9-dimensional binary vectors for training and evaluation.

4. Methodology

4.1. System Overview

Figure 1 illustrates the overall pipeline. Raw tweet text passes through Arabic-specific preprocessing, after which three classifiers produce per-label

probability scores: MARBERT, AraBERT-twitter, and a TF–IDF + Logistic Regression baseline. Their outputs are combined via weighted averaging, and per-label decision thresholds tuned on the validation set convert the ensemble scores into final multi-label predictions.

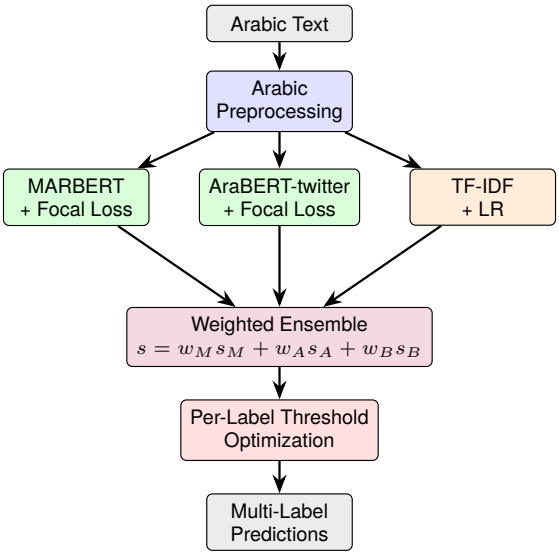

Figure 1: System pipeline for Subtask B multi-label classification.

4.2. Preprocessing

We normalize the tweet text using a light Arabic preprocessing function applied to all splits. The function replaces common character variants (e.g., mapping all forms of alef to ا *alif*, mapping ى *alif maqsura* to ي *ya'*, and normalizing hamza forms), and collapses excessive whitespace. We do not perform aggressive tokenization or stemming, relying instead on the subword tokenizers of the transformer models.

4.3. Handling Class Imbalance

The per-category counts in Table 1 show that Hospitality and Racism / Discrimination are severely under-represented, with only a few instances in the training set. To mitigate this, we perform simple oversampling at the row level. We first identify all texts whose label list includes at least one minority label (Hospitality, Racism / Discrimination, or Greetings), then replicate those rows k times and concatenate them with the original training set. We set $k = 3$, which provides a favorable balance between augmenting minority-label instances and avoiding excessive over-representation relative to the majority classes; preliminary comparisons with $k \in \{2, 5\}$ showed that $k = 3$ maximized validation macro-F1.

Figure 2 shows the category distribution before and after oversampling. The counts for Hospitality and Racism / Discrimination increase substantially, making it more feasible for the models to learn them.

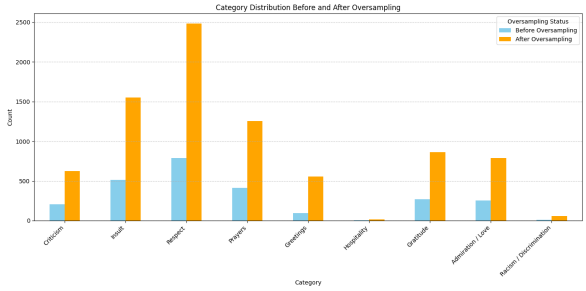

Figure 2: Category distribution before oversampling (blue) and after oversampling (orange).

4.4. Models

We experiment with three models:

- **TF–IDF + Logistic Regression**: A strong lexical baseline using a TfidfVectorizer with up to 20,000 features and character/word n-grams, followed by a One-vs-Rest Logistic Regression classifier.
- **MARBERT**: A transformer encoder pretrained on large-scale Arabic social media, fine-tuned with a 9-dimensional sigmoid output layer for multi-label classification.
- **AraBERT-twitter**: A variant of AraBERT tailored to Twitter data, fine-tuned similarly with a 9-dimensional multi-label head.

For MARBERT and AraBERT-twitter, we set `problem_type="multi_label_classification"` in the Hugging Face implementation and use BCE-with-logits as the base loss.

4.5. Focal Loss

To further address label imbalance, we replace the standard BCE loss with a Focal Loss variant . Given logits z and binary targets $y \in \{0, 1\}$, we first compute the per-label BCE loss and then reweight it as:

$$\text{FL}(z, y) = \alpha(1 - p_t)^{\gamma} \cdot \text{BCE}(z, y),$$

where p_t is the model's probability assigned to the target class, α is a balancing factor, and γ controls the focusing strength. We adopt the standard values $\alpha = 0.25$ and $\gamma = 2.0$ which were shown to

work well across a range of imbalanced classification settings. Focal Loss down weights easy examples and emphasizes hard or misclassified examples, which is desirable in the presence of severe class imbalance.

We implement Focal Loss in a custom Trainer subclass that overrides the default loss computation and otherwise follows standard Hugging Face training.

4.6. Dataset Statistics

After label mapping and oversampling, the final training set contains 6,465 examples (vs. 2,049 original). Table 3 summarizes key statistics:

Metric	Value
Original train size	2,049
Balanced train size	6,465
Validation size	291
Test size	588
Categories	9
Avg labels per tweet	1.75
Minority class boost	$3\times$

Table 3: Dataset statistics after preprocessing.

The validation set remains unchanged (291 examples) to ensure fair threshold tuning.

4.7. Training Setup

We fine-tune MARBERT and AraBERT-twitter for up to 8 epochs with early stopping (patience = 2) based on validation macro-F1. Key hyperparameters include:

- Batch size: 16 (training), 32 (evaluation)
- Learning rate: 2×10^{-5} with cosine decay and 10% warmup
- Weight decay: 0.01
- Max sequence length: 128 tokens
- Mixed precision: fp16 (when GPU available)

For the TF-IDF baseline, we use 20,000 features with unigram+bigram character and word n-grams, and Logistic Regression with 2,000 max iterations per binary classifier. Training takes approximately 30 minutes per model on a single NVIDIA T4 GPU (Google Colab), with MARBERT slightly faster due to its smaller vocabulary.

5. Experiments and Results

5.1. Evaluation Protocol

We treat the validation set as a held-out evaluation set. For each model, we obtain sigmoid probabilities for all nine labels and then convert them into binary predictions using thresholds. Macro-averaged F1 is computed over the nine categories. We also report per-category F1 to highlight the effect on minority labels.

5.2. Single-Model Performance

We first evaluate the TF–IDF baseline and individual transformer models using a single global threshold per model, tuned on the validation set. Table 4 summarizes the macro-F1 scores.

Model	Macro F1
TF–IDF + Logistic Regression	0.3665
MARBERT + Focal Loss	0.2237
AraBERT-twitter + Focal Loss	0.2237

Table 4: Macro-F1 scores on the validation set for single models (global threshold).

The TF–IDF baseline provides a reasonable starting point, and at the global threshold, both transformer models underperform it. Analysis of the per-class outputs reveals that at the optimal global threshold, both MARBERT and AraBERT-twitter tend toward predicting all labels as positive (precision ≈ 0.14, recall $= 1.0$ for most categories), resulting in near-zero F1 for rare classes. We attribute this to the interaction between Focal Loss and severe class imbalance: with very few positive instances per label, the models learn to predict high probabilities broadly in order to minimize focal loss. This degenerate behavior is mitigated in the ensemble through per-label threshold tuning, which allows each label to be assessed at its own calibrated operating point.

5.3. Per-Category F1 Breakdown

Table 5.3 reports per-category F1 scores on the validation set for each individual model and for the final ensemble. The ensemble consistently improves over both transformers and, for most categories, over the TF–IDF baseline as well.

Category	TF-IDF	MARBERT	AraBERT	Ensemble
Criticism	0.00	0.19	0.19	0.37
Insult	0.60	0.41	0.41	0.78
Respect	0.80	0.55	0.55	0.85
Prayers	0.71	0.34	0.34	0.85
Greetings	0.32	0.10	0.10	0.72
Hospitality	0.00	0.01	0.01	1.00
Gratitude	0.51	0.18	0.18	0.70
Admiration / Love	0.36	0.24	0.24	0.60
Racism / Discrim.	0.00	0.01	0.01	0.01
Macro Avg	**0.37**	**0.22**	**0.22**	**0.65**

Table 5: Per-category and macro-averaged F1 on the validation set. Individual transformer scores are reported at their optimal global threshold. Ensemble scores use per-label thresholds.

Per-category and macro-averaged F1 on the validation set. Individual transformer scores are reported at their optimal global threshold. Ensemble scores use per-label thresholds. TF-IDF per-category scores to be populated from final classification report.

5.4. Ensembling and Threshold Optimization

To combine the strengths of all three models, we build a linear ensemble over their probability outputs. We consider weighted combinations of the form:

$$s = w_{\text{M}} s_{\text{MARBERT}} + w_{\text{A}} s_{\text{AraBERT}} + w_{\text{B}} s_{\text{TFIDF}},$$

where the weights sum to 1. For each weight triple, we optimize label-specific thresholds on the validation set by searching over the grid $[0.02, 0.70]$ in steps of 0.01, selecting the threshold for each label independently by maximizing binary F1.

Table 6 reports macro-F1 for all six candidate weight triples evaluated. We select the combination that yields the best validation macro-F1.

MARBERT	AraBERT	TF-IDF	Macro F1
0.40	0.30	0.30	0.6478
0.40	0.40	0.20	0.6463
0.50	0.30	0.20	0.6473
0.30	0.30	0.40	0.6507
0.20	0.30	0.50	0.6523
0.30	**0.20**	**0.50**	**0.6525**

Table 6: Validation macro-F1 for all candidate ensemble weight triples. The best configuration (bold) places the highest weight on the TF-IDF baseline and lower complementary weights on the two transformers.

The final ensemble with weights $(w_M, w_A, w_B) = (0.3, 0.2, 0.5)$ achieves a macro-F1 of 0.6525 on the validation set, substantially outperforming all individual models.

5.5. Per-Category Effects

The ensemble improves F1 not only on frequent categories such as Respect and Insult, but also on minority labels. Qualitatively, MARBERT tends to provide strong representations for noisy social-media text, AraBERT-twitter complements it on some categories, and the TF–IDF baseline remains competitive for lexically distinctive labels such as Insult and Racism / Discrimination. The per-label threshold optimization reduces the tendency of the models to over-predict all labels and yields more calibrated predictions.

5.6. Test Set Evaluation

For the final submission, we apply the trained ensemble with best-found weights $(0.3, 0.2, 0.5)$ and per-label thresholds to the held-out test set. Predictions are generated following the same inference pipeline as for the validation set: text is preprocessed identically, each model produces sigmoid probabilities, scores are combined via the weighted average, and per-label thresholds derived from validation tuning are applied to produce binary predictions. Each text is assigned up to four criteria labels in descending order of ensemble score. Ground-truth labels for the test set are not available to us; final performance achieved a macro F1 of 0.41 on the provided test set.

6. Analysis and Discussion

The experiments highlight several important aspects of this task. First, we noticed that careful label mapping is essential, as many raw annotation strings are inconsistent, and collapsing them into a small, fixed category set improves both label quality and interpretability. Second, even simple oversampling of minority-label tweets significantly helps the models recognize rare categories, particularly when combined with Focal Loss. Third, the ensemble consistently outperforms individual models, showing that traditional lexical features and deep transformers capture complementary signals. In our case, the TF–IDF baseline plays a stronger role than we expected, especially for categories with highly indicative keywords.

7. Conclusion

We presented a multi-label classification system for Arabic politeness criteria in social media. Our approach combines label mapping, oversampling, transformer-based encoders with Focal Loss, and a weighted ensemble with per-label thresholds. The resulting system achieves a validation macro-F1 of 0.6525, substantially improving over individual models and particularly benefiting rare cate-

gories such as Hospitality and Racism / Discrimination. We plan to extend this work by exploring additional Arabic pre-trained models and investigating more principled calibration and threshold-tuning techniques. Future work will build on the AdabEval shared task (Alqifari et al., 2026) and ADAB dataset (Al-Khalifa et al., 2026).

Ethics Statement

This work uses publicly available Arabic social media data annotated for politeness and impoliteness criteria. Such data may contain offensive, hateful, or discriminatory content, especially in categories like Insult and Racism / Discrimination. We use the data solely for research purposes aimed at improving the detection and understanding of harmful or impolite language, and we do not attempt to deanonymize users or link posts to real identities. Models developed in this work should not be deployed without careful assessment of potential biases, including differential error rates across dialects and demographic groups.

Limitations

Our system is trained and evaluated on a relatively small dataset with a skewed label distribution, which may limit generalization to other domains or platforms. We rely on simple oversampling and Focal Loss for class imbalance, but more advanced reweighting or augmentation techniques could further improve minority-label performance. Per-class thresholds are both tuned and evaluated on the same validation set, which introduces a risk of overfitting to the validation distribution; cross-validation for threshold selection is left to future work. Finally, our analysis focuses on macro-F1 and does not fully explore other aspects such as calibration and human interpretability of the predicted criteria.

Bibliographical References

References

2024. *KSII Transactions on Internet and Information Systems*, 18(2).

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.

Mohamed R. Ezzeldin, Gaber Sallam Salem Abdalla, and Abdoulie Faal. 2025. Enhancing arabic sentiment analysis via <scp>marbert</scp> : Domain adaptation with pseudo-labeling and contrastive learning. *Engineering Reports*, 7(12).

Rania Ibrahim Hithnawi, Mohammad M N Hamarsheh, and Mohammed Maree. 2025. Arabert for arabic cyberbullying detection in facebook comments. *Journal of Cybersecurity*, 11(1).

Zejian Liang, Yunxiang Zhao, Haiwen Xu, Hong Huang, and Luning Chen. 2025. A hybrid model integrating roberta, tf-idf, and attention mechanism for medical query intent classification. *Scientific Reports*, 15(1).

Al-Khalifa, Hend and Ghezaiel, Nadia and Bounnit, Maria and Hamed Alhazmi, Hend and Abdullah Alfear, Noof and Fahad Alqifari, Reem and Masoud Almasoud, Ameera and Ahmed Al-Ghamdi, Sharefah. 2026. *ADAB: Arabic Dataset for Automated Politeness Benchmarking - A Large-Scale Resource for Computational Sociopragmatics*. PID https://doi.org/10.48550/arXiv.2602.13870.