SHU at AdabEval 2026: Category-Aware Fine-Tuning of MARBERT for Arabic Politeness and Pragmatic Function Classification

Alla Zwawi, Stephen Wu

Saraya Hamra University
Alsabaa Street, Tripoli, Libya
{a.zwawi, Dr.Stephen.wu}@shu.edu.ly

Abstract

This paper describes our submission to the AdabEval 2026 shared task, addressing Subtask A (politeness classification) and Subtask B (multi-label pragmatic category prediction). For Subtask A, we fine-tuned MARBERT using weighted cross-entropy to mitigate class imbalance. For Subtask B, we apply BCEWithLogitsloss with inverse-frequency positive weighting to address the minority categories, and we introduce a category merging strategy to reduce categories' sparsity and annotation variation. Finally, we propose a stacked architecture where predicted pragmatic categories are injected as auxiliary features into the politeness classifier. Our results demonstrate that dialect-aware modelling, class-imbalance handling, and category-aware stacking improve Macro-F1 across both subtasks, achieving 0.85 for Subtask A and 0.55 for Subtask B on the test set.

Keywords: Politeness Detection, Transformer Encoder Fine-Tuning, Arabic NLP, MARBERT

1. Introduction

Politeness detection in Arabic text presents a challenge due to the language's various linguistic strategies to convey politeness rooted in cultural norms and social hierarchies (Ameri et al., 2023).; social media text is particularly challenging due to dialectal diversity, orthographic variability, sarcasm, and culturally grounded expressions. Additionally, Arabic posts on social media often contain mixtures of Modern Standard Arabic (MSA) and regional dialects (Egyptian, Gulf, Levantine, and Maghrebi), along with informal spelling, elongations, and code-switching. Such characteristics present challenges in the automatic modelling of politeness and pragmatic functions. This shared task (Alqifari et al., 2026) addresses two related problems: Subtask A, which requires assigning one of three politeness labels (Polite, Neutral, Impolite) to each text sample, and Subtask B, a multi-label prediction of nine culturally grounded pragmatic categories (e.g., Criticism, Insult, Respect, Gratitude, Racism/Discrimination).

To motivate base-model selection, we conducted dialect identification using CAMeL Tools (Obeid et al., 2020) and exploratory data analysis to investigate orthography variation, code switching, and elongation. Our analysis revealed a substantial dialect content (over 70%) as well as mixed language text and elongation. Therefore, we selected MARBERT (Abdul-Mageed et al., 2021), a transformer model pretrained on large-scale Arabic Twitter data with broad dialect coverage. Since both subtasks are classification problems, we adopted an encoder-only architecture to enable efficient fine-tuning with lower computational cost.

Our approach combines dialect-aware modelling with class imbalance mitigation. For Subtask A, we apply weighted cross-entropy to address the dominance of the Neutral class. For Subtask B, we formulate the task as multi-label classification using BCEWithLogitsLoss with inverse-frequency positive weighting to improve minority category recall. We further introduce a label-merging strategy to reduce annotation sparsity. Finally, we propose a category-aware stacking method in which predicted pragmatic category probabilities are incorporated into the politeness classifier.

Our experiments show that imbalance-adjusted training improves macro-level performance, that merging semantically overlapping criteria stabilises multi-label prediction, and that injecting pragmatic categories into politeness modelling reduces model confusion. Our approaches ranked 7th for Subtask A with a Macro-F1 of 0.85, and 2nd for Subtask B with a Macro-F1 of 0.55, evaluated on test data. Remaining challenges include sarcasm detection, which contributes to the model confusion between neutral and impolite classes, overlapping pragmatic categories such as Criticism and Insult, and calibration issues for underrepresented categories.

2. Background

The dataset is an annotated corpus consisting of Arabic social media posts from multiple sources (Shein, YouTube, X previously Twitter, etc). Each record contains an identifier, text, source, and primary label for subtask A (Polite, Impolite, Neutral), as well as up to four keyword-category pairs for subtask B. (Al-Khalifa et al., 2026). The input/output format is summarised in Table 1.

Input	Output Subtask A	Output Subtask B
لقاء مفيد وفريد شكرا جزيلا	Polite	Thanks & gratitude - الشكر و الامتنان

Table 1 Example of Input & Output

The dataset for subtask A consists of 4895 records for training, 693 records for evaluation, and 1406 records for testing. The data indicates a clear class imbalance, with Neutral class being the majority class, followed by Polite, while Impolite is the least represented. This motivates

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 174–178
May 11, 2026.

us to use imbalance-mitigation training strategies to ensure improved recall for minority classes.

For Subtask B, the dataset contains 2,049 training records, 291 evaluation records, and 588 test records. The task involves predicting one or more of nine culturally grounded pragmatic categories: Criticism, Insult, Respect, Prayers, Greetings, Hospitality, Gratitude, Admiration / Love, and Racism / Discrimination. Subtask B also exhibits class imbalance, with categories such as Criticism and Respect occurring more frequently, while Hospitality and Racism / Discrimination are extremely rare. Furthermore, the dataset suffers from fragmentation and sparsity due to annotation criteria that are semantically overlapping or presented using different variations, which we addressed by introducing a category merging strategy in the preprocessing stage.

Al-Khalifa et al. (2024) presented a preliminary study on Arabic politeness detection with adequate performance using MARBERT as a base model. Their findings demonstrate the effectiveness of transformer-based transfer learning for modelling politeness in Arabic. However, their study focused on Modern Standard Arabic text and did not explicitly address the challenges posed by dialectal variation, multi-dialect mixing, or pragmatic multi-label classification in noisy social media environments. Our approach addresses these challenges in four ways:

- Dialect-Aware Modelling: Used MARBERT, guided by dialect analysis, for better coverage of dialectal social media text.
- Robust Arabic Text Normalisation: Applied a preprocessing pipeline for noisy dialectal content.
- Label-Merging Strategy (Subtask B): Merged semantically overlapping/variant criteria into nine official categories to reduce sparsity and annotation fragmentation.
- Category-Aware Stacking (Subtask A): Incorporated predicted pragmatic category probabilities from Subtask B into the Subtask A classifier, leveraging the pragmatics-politeness relation.

3. System Overview

Our system is based on fine-tuning MARBERT, a transformer encoder pretrained on a large set of Arabic tweets covering various dialects. We designed a specific model for subtask A and another for subtask B while sharing a unified preprocessing pipeline tailored to social media text noise.

3.1 Text Preprocessing

We implemented a text normalisation pipeline designed to reduce orthographic variation while maintaining sentiment and pragmatic signals with the following steps:

- URL Removal: All hyperlinks (e.g., http..., www...) are removed.
- Hashtag Normalisation: Hashtag symbols (#) are removed while preserving the underlying word.
- Tatweel Removal: The elongation character (ـ) is removed.
- Diacritics Removal: Arabic diacritics are stripped to reduce noise.
- Alef Normalisation: Variants of Alef (آ ,أ ,إ) are normalised to (ا).
- Ya Normalisation: Alif Maqsura (ى) is normalised to (ي).
- Repeated Character Reduction: Sequences of three or more repeated characters are reduced to at most two (e.g., طوووول, طوول).
- Latin Letter Removal: Non-Arabic letters are removed to reduce code-switching noise.
- Emoji Preservation: Emojis are retained, as they encode sentiment and pragmatic cues.
- Whitespace Normalisation: Multiple spaces, tabs, and line breaks are collapsed into a single space.

3.2 Subtask A Model

We fine-tune MARBERT[1] by attaching a linear classification head to the contextualised [CLS] token representation. To mitigate class imbalance, we compute inverse-frequency class weights and apply cross-entropy loss during training.

3.3 Subtask B Model

We fine-tune MARBERT with a sigmoid output layer producing independent probabilities for each category. Training uses BCEWithLogitsLoss with a positive class weighting term computed as:

$$weight_{pos} = \frac{count_{neg}}{count_{pos}}$$

This weighting increases the penalty for false negatives in rare classes, improving recall for infrequent categories such as Hospitality and Racism / Discrimination. At inference time, labels with probability ≥ 0.5 are selected; if none exceed the threshold, the highest-probability category is chosen, and if more than four are predicted, the top four are retained to comply with task constraints.

3.4 Category Merging Rules

For Subtask B, the dataset annotations contain semantically overlapping or non-official categories, which contribute to label sparsity and fragmentation. To reduce this issue, we introduce rule-based merging that maps related variants into the nine official categories. Asking for Permission is mapped to Respect; Threat and Verbal Violence are mapped to Insult; Sarcasm and Accusation variants are mapped to Criticism; and Sectarian or discrimination-related variants

[1] https://huggingface.co/UBC-NLP/MARBERT

are mapped to Racism / Discrimination. This consolidation reduces sparsity and improves label consistency.

3.5 Category-Aware Stacked Model

To utilise the relationship between pragmatic function and politeness, we incorporate Subtask B predictions into Subtask A through a stacked architecture. After training the multi-label subtask B model, we obtain the category probability vector $p_B \in \mathbb{R}^9$ from the model's predicted output distribution. Specifically, for each instance in the Subtask A training data we generate p_B, by applying Subtask B model in inference mode, and concatenate it with the MARBERT [CLS] embedding h used for politeness classification as follows:

$$z = [h \oplus p_B]$$

The combined representation is passed through a projection and classification layer to produce the final politeness prediction. This integration allows pragmatic signals, such as Insult and Gratitude, to directly inform politeness modelling, in aim to reduce confusion between classes.

4. Experimental Setup

We use the official train/eval/test splits provided by the shared task organisers. We use the training and evaluation sets for model training and hyperparameter validation, while final predictions are generated on the held-out test sets.

All experiments use the preprocessing pipeline described in Section 3.1. Tokenisation is performed using the MARBERT tokeniser with a maximum sequence length of 128 tokens, and Inputs longer than this limit are truncated.

Models are fine-tuned using the AdamW optimiser with a learning rate of 2e-5, a batch size of 16, and 3 epochs of training. Mixed-precision (FP16) training is enabled to improve computational efficiency. For Subtask A, we use weighted cross-entropy loss to address class imbalance. For Subtask B, we use BCEWithLogitsLoss with positive class weights ($weight_{pos}$).

Our implementation uses the Hugging Face Transformers library (v4.x)[2], PyTorch (v2.x,)[3] and Datasets (v2.x)[4]. Dialect analysis is performed using CAMeL Tools (Obeid et al., 2020)[5].

For Subtask A, we evaluate Accuracy, Macro-Precision, Macro-Recall, and Macro-F1, with Macro-F1 serving as the primary evaluation metric due to class imbalance. For Subtask B, we evaluate Micro-F1, Macro-F1, Macro-Precision, and Macro-Recall. Macro-F1 is emphasised to ensure balanced evaluation across rare and frequent categories. Macro-F1 score on the evaluation set is used for model selection, with the highest-scoring checkpoint chosen as the final submission for official ranking in both subtasks.

5. Results & Discussion

5.1 Quantitative Results

This section details the results of our experiments; Table 2 illustrates the performance of different configurations for Subtask A, evaluated on the validation set. We compare a baseline MARBERT model, a version incorporating text preprocessing and weighted cross-entropy loss, and a category-aware stacked model that integrates Subtask B predictions, with preprocessing and weighted cross-entropy loss.

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
MARBERT (Baseline)	0.91	0.852	0.83	0.839
MARBERT +Preproces +CrossEntropy	0.91	0.854	**0.838**	**0.843**
MARBERT Category-Aware	**0.913**	**0.856**	0.827	0.839

Table 2 Subtask A Performance Evaluation

The baseline MARBERT model already achieves strong performance due to its pretraining on Arabic social media data. Incorporating the preprocessing pipeline and class-weighted cross-entropy improves Macro-F1, primarily by increasing recall for minority classes. These improvements confirm that normalisation tailored to dialectal Arabic text reduces orthographic noise and improves model generalisation.

The category-aware stacked model achieves the highest accuracy (0.913) and macro-precision (0.856). However, its Macro-F1 remains comparable to the baseline because the additional pragmatic features slightly reduce macro-recall. This suggests that pragmatic signals may improve precision in Arabic politeness detection but may also lead to confusion in borderline cases.

Model	Micro-F1	Macro-Precision	Macro-Recall	Macro-F1
MARBERT +BCEWithLogitsLoss (Baseline)	0.773	0.669	0.531	0.58

[2] https://huggingface.co/transformers
[3] https://pytorch.org
[4] https://huggingface.co/docs/datasets
[5] https://camel-tools.readthedocs.io

MARBERT +BCEWith LogitsLoss +Criteria Merging	**0.784**	**0.669**	**0.922**	**0.746**

Table 3 Subtask B Performance Evaluation

Table 3 presents the evaluation results for Subtask B, comparing the baseline MARBERT multi-label classifier, fine-tuned with BCEWithLogitsLoss, and the configuration incorporating a criteria-merging strategy applied as a preprocessing step. This strategy consolidates semantically overlapping labels prior to model training. The baseline model achieves a Micro-F1 of 0.773 but a relatively low Macro-F1 of 0.580, reflecting the strong class imbalance in the dataset, where frequent categories dominate the micro-averaged metric while rare categories remain poorly predicted.

Introducing the criteria-merging strategy significantly improves Macro-F1, mainly due to an increase in macro-recall. The criteria-merging strategy increases effective training instances for sparse labels and reduces fragmentation, advancing the model's ability to generalise across pragmatically related expressions. The Micro-F1 also increases slightly, indicating that the merging strategy improves performance without sacrificing accuracy on frequent categories.

5.2 Error Analysis

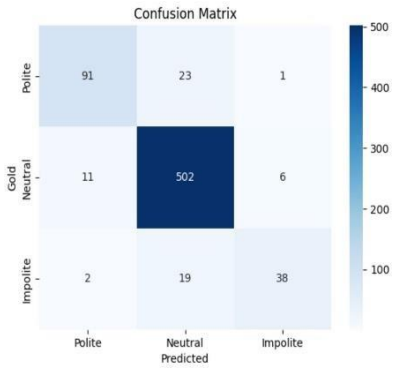

Figure 1 Confusion Matrix for Subtask A

For Subtask A, we analyse model errors using the confusion matrix shown in Figure 1 and qualitative analysis. The model performs best on the Neutral class, reflecting the model's bias toward the majority class. The Polite class is predicted reliably, although 20% Polite instances are misclassified as Neutral when explicit politeness markers (e.g., gratitude, greetings) are absent. The most challenging class is Impolite, where only 64% of instances are classified correctly, compared to 79% for Polite and 97% for Neutral. 32% of impolite instances are misclassified as Neutral which suggests that negative or critical expressions are often interpreted as neutral when explicit insult cues are not present.

Qualitative inspection reveals recurring error types, supported by quantitative trends in misclassified Impolite instances. Among these errors, 35% contain evaluative words such as "سيء" (bad), where negative judgment is expressed without explicit offense, leading to Neutral predictions. Sarcasm accounts for 25% of errors, reflecting the model's reliance on lexical cues rather than intended meaning. Similarly, 20% involve religious expressions (e.g., " والله ان تربيتكم صحيحه") used to convey indirect criticism, which are often interpreted literally. These patterns align with confusion matrix results, where Impolite instances are frequently misclassified as Neutral, indicating systematic difficulty in capturing implicit negativity. Notably, many misclassified examples have high prediction confidence (90% or higher), indicating the model is confidently incorrect when pragmatic meaning differs from lexical sentiment.

Error analysis for Subtask B shows that errors are concentrated in rare categories, particularly Hospitality and Racism/Discrimination (support ≈1), reflecting severe class imbalance which leads the model to default to more frequent categories, such as Criticism or Respect, instead. Additionally, semantic overlap between pragmatic categories (e.g., admiration, gratitude, respect) which appear in similar linguistic contexts leads to label confusion. These results highlight class imbalance and pragmatic overlap as key challenges in multi-label politeness classification for Arabic social media posts.

6. Conclusion

This paper presented a system for Arabic politeness detection and pragmatic category prediction in social media text. We fine-tuned MARBERT for both subtasks and introduced improvements, including Arabic text normalisation, class-imbalance handling, a label-merging strategy for Subtask B, and a category-aware stacked model integrating pragmatic signals into politeness classification. Results show that MARBERT performs strongly across subtasks; while preprocessing and label consolidation improve robustness under dialectal variation and sparse annotations. Despite these contributions, several limitations remain. The approach relies on predicted pragmatic labels, which may propagate errors in the stacked model. Performance is also constrained by severe class imbalance and limited support for rare categories, and difficulty capturing implicit meaning such as sarcasm and indirect criticism. Future work should focus on improving robustness to implicit pragmatic cues, particularly sarcasm, and exploring more reliable integration of pragmatic signals.

7. Bibliographic References

Ameri, M., Zeighami, A., & Mirahmadi, S. R. (2023).A study of the polite method in the Arabic language according to Brown and Levinson's theory of politeness. Studies on Arabic Language and Literature,13 (36), 61–88. https://doi.org/10.22075/lasem.2022.23600.1286

Hend Al-Khalifa, Nadia Ghezaiel, and Maria Bounnit. 2024. Analyzing Politeness in Arabic Tweets: A Preliminary Study. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), pages 352–359, Trento. 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 Ahmed Al-Ghamdi. ADAB: Arabic Dataset for Automated Politeness Benchmarking - A Large-Scale Resource for Computational Sociopragmatics. 2026. In Proceedings of The 2026 International Conference on Language Resources and Evaluation (LREC2026), Palma, Mallorca (Spain).

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.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 7022–7032, Marseille, France. European Language Resources Association.

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).