The AdabEval 2026 Shared Task on Arabic Politeness Detection

Reem Alqifari[1], Hend Al-Khalifa[1], Nadia Ghezaiel Hammouda[2], Maria Bounnit[3]
Hend AlHazmi[4], Ameera Almasoud[1], Sharefah Al-Ghamdi[1], Noof Alfear[1]
[1]College of Computer and Information Sciences, King Saud University,
[2]College of Computer Science and Engineering, University of Hail,
[3]Cadi Ayyad University, [4]Saudi Center of Philosophy and Ethics
[1]Riyadh, Saudi Arabia, [2]Hail, Saudi Arabia,
[3]Marrakesh, Morocco, [4]Jeddah, Saudi Arabia
[1]ralgifary|hendk|ammalmasoud|sharefah|nalfear@ksu.edu.sa,
[2]ghezaielnadia.ing|[3]mariabounnit|[4]hend.hamed.w@gmail.com

Abstract

We present an overview of the AdabEval 2026 shared task, organized as part of the OSACT7 workshop (co-located with LREC 2026). This task introduces the first benchmark suite for politeness detection. It includes two subtasks: Politeness Classification (Subtask A) and Category Prediction (Subtask B). The task focuses on evaluating models' ability to recognize and categorize politeness phenomena in Arabic text. Evaluation was conducted using an automatic metric (macro F1-score). A total of 28 unique teams participated in the shared task. Of these, 13 teams submitted final system predictions across the two subtasks. The top-performing systems relied primarily on transformer-based architectures. The winning systems achieved macro F1-scores of 0.89 for Subtask A and 0.58 for Subtask B.

Keywords: Politeness Classification, Politeness Category Prediction, Arabic Natural Language Processing

1. Introduction

Politeness plays a central role in conversation, languages provide diverse ways to encode respect and reduce imposition, and these markers are intimately tied to social power and interpersonal harmony. In Arabic, the concept is deeply rooted in cultural norms; speakers employ honorifics, plural forms of address and kinship terms to show respect. Most computational work on politeness focuses on English, with only limited studies on other languages. The ADAB dataset introduced the first large-scale Arabic politeness dataset, containing 10,000 annotated examples across three politeness categories (Al-Khalifa et al., 2026). The dataset used in this shared task is derived as a subset of ADAB. To the best of our knowledge, this shared task represents the first Arabic benchmark focused on politeness analysis, addressing both politeness prediction and the identification of politeness categories.

The shared task consists of two complementary subtasks: **Subtask A: Politeness Classification**, which focuses on identifying the overall politeness level of a text, and **Subtask B: Politeness Category Prediction**, which aims to identify the underlying pragmatic strategies expressed in the text. This distinction reflects the fact that politeness involves not only surface-level classification but also multiple interacting communicative functions within the same utterance.

Modeling politeness in Arabic presents several challenges. Politeness strategies are often expressed implicitly and may rely on cultural conventions, formulaic expressions, or contextual cues. In addition, multiple categories may co-occur within a single sentence, making fine-grained prediction particularly challenging.

This overview paper describes the design and organization of the AdabEval shared tasks. We present the task definitions, dataset, evaluation protocol, baseline systems, and a summary of participants results. We conclude with insights gained from the shared task and directions for future research.

2. Task Description

The shared task consists of two subtasks. Subtask A focuses on building and evaluating models that automatically assess the politeness level of Arabic text. Given a text, participants must classify it into one of three categories: Polite, Neutral, or Impolite. Systems will be compared using macro-averaged F1-score on the test data. Subtask B evaluates the ability of systems to identify multiple pragmatic functions in Arabic social-media posts. Each text may express one or more categories from nine culturally grounded functions including: "Criticism", "Insult", "Respect", "Prayers", "Greetings", "Hospitality", "Gratitude", "Admiration_Love", and "Racism_Discrimination". The task is framed as multi-label classification, and systems must predict all applicable categories for each instance. The of-

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 151–156
May 11, 2026.

Task	Train	Val	Test
Subtask A	4896	694	1407
Subtask B	2050	292	589

Table 1: Dataset split sizes for the AdabEval shared tasks.

ficial evaluation metric is macro-averaged F1-score across the nine categories.

3. Dataset Description

The dataset used in this shared task is derived as a subset of the ADAB dataset (Al-Khalifa et al., 2026), a large-scale Arabic politeness corpus. The subset was constructed to support the two subtasks, focusing on politeness classification and fine-grained politeness category prediction. Both subtasks rely on sentences collected from multiple online platforms and annotated according to a unified annotation scheme. Each sentence in the dataset is associated with a source platform indicating where the text was originally collected. The sources include YouTube comments, Shein product reviews, Twitter posts, and company product reviews. This diversity of sources allows the dataset to capture different styles of user-generated Arabic text across multiple communication contexts, including the natural mixture of Modern Standard Arabic (MSA) and dialectal Arabic commonly found in online communication.

3.1. Subtask A: Politeness Classification

Subtask A focuses on classifying sentences into three politeness classes: *Polite*, *Impolite*, and *Neutral*. Each instance in this task contains three fields: the sentence text, the source platform, and the corresponding politeness label.

The dataset is divided into training, validation, and test splits. The training set contains 4,896 instances, the validation set contains 694 instances, and the test set contains 1,407 instances. To ensure fair evaluation, the test set provided to participants includes only the sentence text and source information, and participants are required to predict the politeness label.

Table 1 summarizes the dataset splits for Subtask A, while the label distribution is illustrated in Figure 1.

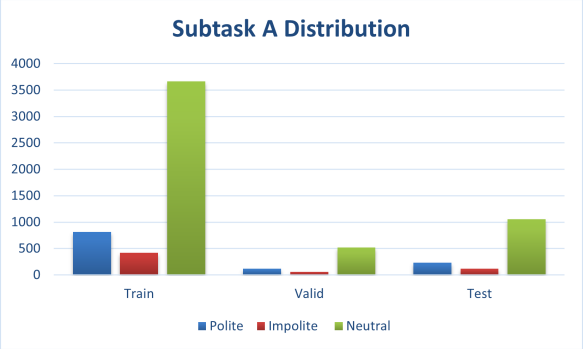

Figure 1: Distribution of politeness labels in the Subtask A

3.2. Subtask B: Category Prediction

Subtask B focuses on identifying the politeness category expressed in sentences labeled as either *Polite* or *Impolite*. For this task, a subset of the original dataset was selected that contains only sentences belonging to these two classes.

Each instance in the training and validation sets includes the following fields: the sentence text, the source platform, the politeness label, and up to four annotated politeness criteria. Each criterion is accompanied by a keyword that represents the lexical trigger associated with the identified strategy. These annotations were originally introduced to capture pragmatic signals underlying politeness expressions. For consistency in formatting, the dataset is represented using four columns (criteria1–criteria4), where unused fields are left empty if fewer than four criteria apply.

The dataset is split into training, validation, and test sets. The training set contains 2,050 instances, the validation set contains 292 instances, and the test set contains 589 instances. Similar to Subtask A, the test categories are hidden from participants and only the sentence text, source and label information are provided. Table 1 summarizes the dataset splits for Subtask B.

In addition to the binary politeness labels, each sentence in Subtask B is associated with a fine-grained politeness strategy category. These categories capture different pragmatic functions such as respect, criticism, gratitude, and prayers. Table 2 and Figure 2 present the distribution of politeness categories across the training, validation, and test splits. The dataset contains 2,904 instances distributed across several polite and impolite strategies. Overall, polite strategies are more frequent than impolite ones. The most frequent category is Respect (1,034 instances), followed by Criticism (637 instances). The category Prayers appears in both the polite and impolite classes. Additionally, some categories such as Racism_Discrimination and Hospitality, contain no instances in the current dataset due to the natural distribution of the collected data.

Category	Politeness	Train	Valid	Test
Criticism	Impolite	435	78	124
Insult	Impolite	155	22	64
Racism_Discrimination	Impolite	0	0	0
Prayers	Impolite	58	2	12
Prayers	Polite	357	57	95
Respect	Polite	741	105	188
Greetings	Polite	79	16	17
Hospitality	Polite	0	0	0
Gratitude	Polite	168	16	43
Admiration_Love	Polite	43	8	21

Table 2: Distribution of politeness categories across dataset splits.

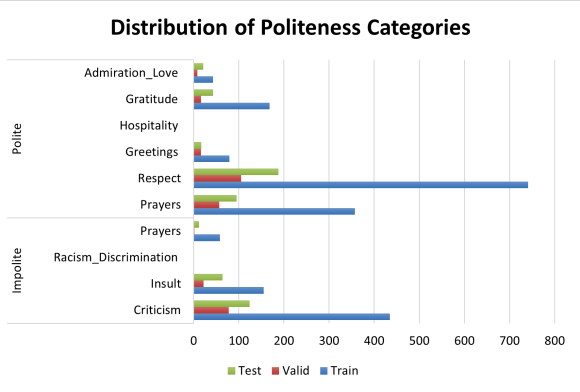

Figure 2: Distribution of politeness categories across dataset splits in Subtask B.

4. Evaluation

This section describes the evaluation protocol and submission procedures used in the AdabEval shared tasks. The evaluation was conducted using the Codabench platform[1], which allows participants to submit their system predictions and automatically evaluates them against the hidden test labels.

4.1. Evaluation Metric

For both subtasks, system performance is evaluated using the macro-averaged F1 score. This metric computes the F1 score independently for each class and then averages the scores across all classes. Macro-F1 is particularly suitable for this task because it assigns equal importance to all classes regardless of their frequency in the dataset, thereby mitigating the effects of class imbalance.

For Subtask A, the evaluation considers three classes: *Polite*, *Neutral*, and *Impolite*. For Subtask B, the evaluation measures the ability of systems to correctly predict the politeness category associated with each sentence.

[1] https://www.codabench.org/

4.2. Submission Format

Participants were required to submit their predictions through the Codabench platform as a single compressed ZIP file. The ZIP archive must contain the prediction files corresponding to the subtasks.

Subtask A. For Subtask A, the ZIP file must include a file named `subtask_A.csv`. The CSV file must contain exactly one column named `label`, where each row corresponds to a predicted politeness label for a test instance. The predicted label must be one of the following classes: *Polite*, *Neutral*, or *Impolite*. Each test instance must receive exactly one prediction, and the row order in the submission file must match the order of the instances in the provided test set.

Subtask B. For Subtask B, the ZIP file must include a file named `subtask_B.csv`. The CSV file must contain exactly four columns named `criteria1`, `criteria2`, `criteria3`, and `criteria4`. Each column contains a predicted politeness strategy category, or is left empty if fewer than four categories apply. Each sentence may be assigned one or more categories, with a maximum of four categories per instance. Although each instance may have fewer than four labels, a fixed four-column format is used to standardize submissions.

Each row in the file corresponds to a test instance, and the row order must exactly match the order of the provided test data. Category labels are not mutually exclusive and must be selected from the official list of Subtask B categories. Submissions that do not follow the specified format may be rejected or receive a score of zero.

4.3. Evaluation Procedure

All submissions were evaluated using the Codabench platform. We implemented a custom scoring script that automatically computes the evaluation metric for each submitted prediction file. The scoring script reads the submitted files, validates their format, and compares the predictions against the hidden ground-truth labels of the test set.

For Subtask A, the script computes the macro-averaged F1 score over the three politeness classes. For Subtask B, the script evaluates the predicted politeness categories for each sentence and calculates the macro-averaged F1 score across all categories.

The scoring script also ensures that the number and order of predictions match the provided test data. Submissions that do not follow the required format are rejected or receive a score of zero.

Task	Val Macro-F1	Test Macro-F1
Subtask A	0.83	0.85
Subtask B	0.41	0.38

Table 3: Performance of the MARBERT baseline models.

4.4. Competition Phases

The shared task was organized into two phases to facilitate system development and fair evaluation. During the evaluation phase, participants were provided with the training and validation datasets to train and tune their models. Participants could submit predictions to the Codabench leaderboard and receive evaluation scores based on the validation set. In the final test phase, participants were required to submit their predictions for the hidden test set. The ground-truth labels for the test data were not released, and the final ranking of participating systems was determined based on their performance on the test set.

5. Baseline Systems

To provide a reference point for participating teams, we implemented baseline systems for both subtasks using a transformer-based Arabic language model. Specifically, we fine-tuned MARBERT (Abdul-Mageed et al., 2021), a pre-trained language model designed for Arabic social media text, on the training data of each subtask. MARBERT was selected as the baseline model due to its strong performance reported in prior work comparing a wide range of approaches, including traditional machine learning models, transformer-based models, and large language models (Al-Khalifa et al., 2026).

For Subtask A, the model was fine-tuned to perform multi-class classification over the three politeness labels. The model takes the input sentence and predicts one of the three classes.

For Subtask B, MARBERT was adapted to predict the politeness categories associated with each sentence. Since multiple categories may apply to a single sentence, the model was trained in a multi-label classification setting, allowing it to assign one or more strategy labels per instance.

The baseline models were trained using the official training splits provided for each task, and performance was evaluated on both the validation and test sets using the same evaluation metric described in Section 4.1. Table 3 reports the performance of the baseline systems for both subtasks.

6. Participating Systems

The AdabEval shared tasks attracted participation from multiple research teams working on Arabic NLP. Different teams participated in each subtask, reflecting varying research interests in politeness classification and politeness category prediction.

A total of 28 unique teams registered for the AdabEval shared tasks across both subtasks.

For Subtask A, 19 teams registered, of which 8 submitted final system predictions. For Subtask B, 13 teams registered, of which 5 submitted final results.

These numbers are not mutually exclusive, as some teams participated in both subtasks. Participants were allowed to compete in one or both subtasks independently.

The majority of participating systems relied on transformer-based architectures, particularly Arabic pre-trained language models such as AraBERT(Antoun et al., 2020) and MARBERT. Some teams also explored multilingual transformer models and ensemble methods.

Detailed descriptions of the individual systems are provided in the corresponding system papers submitted by the participating teams. Several teams submitted system predictions but did not submit system description papers. Five teams submitted system description papers. In the next section of this overview paper, we report the official results of the shared tasks based on the evaluation of system predictions on the hidden test sets.

Rank	**Team**	**Macro-F1**
1	grkurdi*	0.89
2	Nina*	0.87
3	abdllh	0.87
4	rand_at	0.85
5	Baseline	0.85

Table 4: Leaderboard results for Subtask A. Teams marked with * submitted system description papers.

7. Results

This section presents the official leaderboard results for the AdabEval shared tasks. System performance was evaluated on the hidden test sets using the macro-averaged F1 score. Tables 4 and 5 report the results for Subtask A and Subtask B, respectively. Teams marked with an asterisk (*) submitted system description papers. The leaderboard tables report only systems that outperform the baseline model.

8. Discussion

The results provide several insights into the challenges of automatic politeness detection in Arabic. Overall, systems achieved stronger performance on Subtask A compared to Subtask B, reflecting

Rank	Team	Macro-F1
1	gfattani*	0.58
2	alla_zwawi*	0.55
3	jana-sami	0.55
4	rand_at*	0.41
5	Baseline	0.38

Table 5: Leaderboard results for Subtask B. Teams marked with * submitted system description papers.

the increased difficulty of predicting fine-grained politeness strategies.

For Subtask A, the results reveal a strong bias toward the neutral class. As the dominant category in the dataset, neutral instances were frequently predicted by the systems, leading to many polite and impolite utterances being misclassified as neutral. This bias suggests that models tend to favor the majority class when explicit politeness cues are absent, highlighting the difficulty of distinguishing subtle politeness signals from neutral expressions.

In Subtask B, performance varied across different categories. This variability can be attributed to several interrelated factors. Categories such as Respect and Prayers proved relatively straightforward to detect. This is largely because such expressions tend to follow recognizable formulaic patterns deeply embedded in Arabic communicative norms. Speakers of Arabic frequently employ fixed religious or honorific phrases in these contexts, providing models with reliable lexical anchors. In contrast, categories such as Criticism and Insult were more challenging because they often rely on contextual or implicit cues.

Furthermore, expressions categorized as Prayers can convey either positive or negative intent depending on context. Prayer expressions may function as polite wishes directed toward someone's well-being (e.g., wishing someone success or good fortune) or as negative imprecations directed toward someone (e.g., wishing harm). Such contextual variation increases the ambiguity of the category and further complicates fine-grained politeness strategy prediction. The multi-label nature of Subtask B added yet another layer of complexity. Unlike single-label classification, where a model selects the most fitting category, multi-label settings require the simultaneous identification of potentially overlapping labels, demanding a more nuanced understanding of the input. Further contributing to lower performance was the issue of class imbalance. When certain categories appear only rarely in the training data, models struggle to learn meaningful representations for them, and this weakness persists even when compensatory techniques such as focal loss or class-weighted training are employed. This was evident in the notably lower scores observed for less frequent categories, where even well-tuned systems failed to achieve reliable detection.

Across both subtasks, the participating systems converged on a broadly similar approach, relying on transformer-based architectures pretrained on Arabic text, particularly AraBERT and MARBERT. These models were fine-tuned using standard classification objectives, with some teams introducing modifications to handle class imbalance more effectively. The strong results observed in Subtask A affirm that pretrained models are well suited to capturing the more explicit and conventionalized markers of politeness in Arabic. However, the comparatively lower performance in Subtask B suggests that these models, despite their considerable capacity, still fall short when confronted with pragmatically subtle or culturally specific expressions.

These observations highlight the importance of contextual and pragmatic understanding when modeling politeness in Arabic. Politeness strategies are often conveyed through culturally specific expressions and indirect formulations that may not be captured by surface lexical cues alone.

9. Conclusion

This paper presented an overview of the AdabEval shared tasks, which aim to advance research on politeness understanding in Arabic. The shared task introduced two complementary subtasks: Subtask A focused on politeness classification into *Polite*, *Neutral*, and *Impolite*, while Subtask B addressed the prediction of fine-grained politeness categories. Both tasks were derived from a curated subset of a previously published Arabic politeness dataset, covering multiple sources of user-generated content.

The shared tasks attracted participation from multiple research teams, who explored a variety of modeling approaches, primarily based on transformer architectures and Arabic pre-trained language models. The leaderboard results demonstrate the progress made in modeling politeness in Arabic, while also highlighting the challenges associated with detecting subtle pragmatic cues and fine-grained politeness categories.

We hope that the AdabEval shared tasks will encourage further research on pragmatic and sociolinguistic aspects of Arabic NLP.

10. Limitations

Despite its contributions, the dataset and tasks have several limitations. First, the data are collected from a limited set of online platforms, which may not fully capture the diversity of Arabic language use across different domains, regions, and dialects.

Second, the dataset exhibits class imbalance. In Subtask A, the *Neutral* class constitutes the majority of instances, which may bias models toward predicting neutral labels. In Subtask B, certain categories (e.g., *Respect* and *Criticism*) are more frequent than others, leading to uneven category representation.

Third, some categories in Subtask B are underrepresented or absent in the training data. For example, categories such as *Hospitality* and *Racism_Discrimination* contain no instances in certain splits, which prevents models from learning reliable patterns for these categories.

Finally, the task is limited to sentence-level classification and does not incorporate broader discourse context, which may be necessary for accurately interpreting politeness in real-world communication.

Future work will aim to address these limitations by expanding the dataset to include more diverse sources, and exploring context-aware approaches to politeness modeling.

11. Ethics Statement

This shared task complies with the LREC 2026 Ethical Guidelines for responsible language resource development. It supports research on politeness understanding in Arabic while adhering to responsible data usage practices. The dataset is derived from publicly available user-generated content and has been anonymized to protect user privacy. No personally identifiable information was intentionally included.

All evaluations were conducted on hidden test labels, which were not accessible to participants, ensuring a fair and unbiased comparison across systems.

12. 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.

Hend Al-Khalifa, Nadia Ghezaiel, Maria Bounnit, Hend Hamed Alhazmi, Noof Abdullah Alfear, Reem Fahad Alqifari, Ameera Masoud Almasoud, and Sharefah Ahmed Al-Ghamdi. 2026. Adab: Arabic dataset for automated politeness benchmarking–a large-scale resource for computational sociopragmatics. *Proceedings of 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, with a Shared Task on Offensive Language Detection*, pages 9–15. European Language Resources Association. Marseille, France.