GHAD NLP at AdabEval2026: Transformer-Based Approach for Arabic Politeness and Pragmatic Category Classification

Ghader Kurdi and Ghada Alfattni
Department of Data Science, Department of Computer Science
{College of Computing, Jamoum University College}, Umm Al-Qura University, Makkah, Saudi Arabia
{grkurdi, gafattni}@uqu.edu.sa

Abstract

This paper presents our submission to the AdabEval 2026 shared task on Arabic politeness classification and pragmatic category prediction. We explored a range of Arabic-specific and multilingual transformer models and integrated their outputs through an ensemble strategy. Our approach achieved state-of-the-art performance in the shared task, ranking first in both subtasks with a macro-F1 score of 0.89 and an accuracy of 0.93 on subtask A, and a macro-F1 score of 0.58 on subtask B. Although our approach delivered high performance on overall politeness classification, pragmatic category prediction remains more challenging. Despite achieving the top ranking in this subtask, the comparatively lower macro-F1 score suggests that modelling fine-grained pragmatic functions requires further methodological refinement and experimentation.

Keywords: Politeness, Classification, Transformers

1. Introduction

Politeness is a fundamental aspect of human communication, shaping how speakers express respect, mitigate imposition, and maintain social harmony. In Arabic, politeness is deeply embedded in cultural and linguistic norms, where speakers employ honorifics, plural forms of address, kinship terms, prayers, and respectful expressions to convey social relationships and interpersonal attitudes. Automatic detection of politeness in Arabic social media is important for various applications such as content moderation and generation, sentiment analysis, dialogue systems, and conversational AI. However, most of the computational work on politeness has focused on English, with limited resources and research available for Arabic (Priya et al., 2024).

The AdabEval 2026 shared task (Alqifari et al., 2026) addresses this gap by introducing a large-scale annotated dataset of Arabic social media posts labelled for politeness and pragmatic functions, enabling systematic evaluation and comparison of computational approaches. The shared task comprises two subtasks: (A) politeness classification, which involves classifying text into polite, neutral, or impolite categories, and (B) category prediction, which requires identifying one or more pragmatic functions, such as criticism, insult, respect, prayer, or greeting. These tasks present unique challenges due to the diversity of Arabic dialects, informal writing styles, the brevity and limited contextual information characteristic of social media texts, and culturally grounded expressions of politeness. In this paper, we describe the GHAD (غد) NLP team's submission to the AdabEval 2026 shared task. Our approach leverages pretrained Arabic and multilingual transformer models fine-tuned for politeness classification and combines them using an ensemble strategy. Pretrained language models capture rich contextual and semantic information, making them particularly effective for pragmatic and sociolinguistic tasks. We explore multiple system configurations and investigate the impact of preprocessing strategies, training objectives, and model architectures. Our approach ranked first in both subtasks, attaining a macro-F1 score of 0.89 and an accuracy of 0.93 on subtask A, and a macro-F1 score of 0.58 on subtask B, highlighting its effectiveness.

Our contributions can be summarised as follows:

- we propose a transformer-based ensemble for Arabic politeness classification;
- we design a MARBERT-based multi-label framework with label-specific threshold optimisation;
- we conduct extensive experiments demonstrating state-of-the-art performance on both subtasks.

2. Background

2.1. Task Definition

The AdabEval 2026 shared task focuses on automatic detection of politeness and pragmatic functions in Arabic social media posts. The challenge consists of two subtasks, each addressing a different dimension of the classification challenge.

Subtask A: Politeness Classification. This subtask is formulated as a multi-class classification

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 157–164
May 11, 2026.

problem. Given an Arabic text, the system must classify it into one of three politeness levels: *Polite*, *Neutral*, or *Impolite*. The goal is to capture the overall politeness expressed in the text.

Subtask B: Category Prediction. This subtask is formulated as a multi-label classification problem. Each text may express one or more pragmatic functions reflecting culturally grounded politeness strategies. The goal is to predict all applicable categories from a predefined set of nine categories: *Criticism*, *Insult*, *Disparagement*, *Prayers*, *Greetings*, *Admiration*, *Respect*, *Felicitation*, and *Hospitality & Generosity*. Examples of polite expressions include prayers and asking for permission, while impolite expressions include insults or criticism.

2.2. Dataset Description

The dataset, released as part of the AdabEval 2026 shared task and made available on Codabench,[1] consists of Arabic social media posts collected from four sources: Companies platforms, Shein, YouTube, and X. More details about the data sources as well as the collection and annotation processes can be found in (Al-Khalifa et al., 2026a). The dataset contains a total of 7,928 instances distributed across multiple files corresponding to different subtasks and data splits (Table 1).

For subtask A (Politeness Classification), the dataset includes a training set containing 4,895 instances and a validation set containing 693 instances. For subtask B (Category Prediction), additional training and validation files focusing on polite and impolite instances were provided, containing 2,049 and 291 instances, respectively.

Each training and validation file contains three columns: *Sentence*, which represents the input text of Arabic social media posts; *Source*, indicating the platform from which the text was collected (Companies, Shein, YouTube, or Tweet); and *label*, corresponding to the ground-truth annotation of the politeness level (Polite, Neutral, or Impolite).

For subtask B, the dataset provides eight additional columns that encode pragmatic annotations. These columns are grouped into four pairs, each consisting of a category label and its corresponding supporting keyword(s). The fields *criteria1–criteria4* indicate the assigned pragmatic categories, while *keywords1–keywords4* provide lexical evidence supporting each category assignment. This design enables multi-label annotation, where a single text may be associated with up to four pragmatic categories.

After the development phase, the test sets were released for final evaluation. The subtask A test set contains 1,406 instances with two columns (*Sentence* and *Source*), with ground-truth labels withheld for blind evaluation via the leaderboard. The subtask B test set includes 588 instances with three columns (*Sentence*, *Source*, and overall *label*), while the fine-grained category labels were withheld to ensure blind evaluation.

The dataset reflects informal Arabic language usage across different social media platforms and includes modern standard Arabic and dialectal Arabic. The diversity of sources introduces variation in linguistic style, politeness strategies, and pragmatic expressions, making the task particularly challenging. Furthermore, linguistic phenomena found in social media posts such as slang usage, short texts, emoji use, and sarcasm pose additional challenges for accurate classification.

2.3. Task Challenges

Politeness detection in Arabic presents several challenges. First, Arabic exhibits rich morphological variation, which increases lexical diversity and complicates modelling. Second, social media texts often contain dialectal variations, spelling inconsistencies, and informal expressions. Third, politeness is a pragmatic phenomenon that depends on cultural norms and contextual interpretation, making it more difficult to capture using purely lexical features. Finally, some texts may express multiple pragmatic functions simultaneously, requiring models capable of handling multi-label classification.

2.4. Related Work

Politeness detection has been widely studied in computational linguistics, primarily focusing on English (Al-Khalifa et al., 2026a). Early work relied on handcrafted linguistic features, including politeness markers, lexical cues, and syntactic structures, to identify politeness strategies (Danescu-Niculescu-Mizil et al., 2013). These approaches were limited in their ability to generalise across domains.

Recent advances in pretrained language models have significantly improved performance across various NLP tasks. BERT (Devlin et al., 2019) and its variants have achieved state-of-the-art results in text classification, sentiment analysis, and pragmatic inference. These models learn contextualised representations that capture semantic and syntactic information, making them well-suited for politeness detection.

For Arabic, several pretrained transformer models have been developed to address the linguistic complexity and morphological richness of the language. AraBERT (Antoun et al., 2020) was one of the first pretrained BERT models specifically designed for Arabic. MARBERT (Abdul-Mageed et al., 2021) further improved performance by training on

[1] https://www.codabench.org/competitions/11955/

Subtask	Split	Neutral	Polite	Impolite
A	Train	3,664 (74.85%)	815 (16.65%)	416 (8.50%)
	Valid	519 (74.89%)	115 (16.59%)	59 (8.51%)
	Total	4,183 (74.86%)	930 (16.64%)	475 (8.50%)
B	Train	-	1,338 (65.30%)	711 (34.70%)
	Valid	-	190 (65.29%)	101 (34.71%)
	Total	-	1,528 (65.30%)	812 (34.70%)

Table 1: Distribution of politeness labels across the training and validation splits.

large-scale Arabic social media data, making it particularly effective for informal and dialectal Arabic. CAMeLBERT (Inoue et al., 2021) introduced a suite of pretrained models trained on diverse Arabic corpora, demonstrating strong performance across multiple downstream tasks.

Despite these advances, politeness detection in Arabic remains relatively under-explored due to the lack of large annotated datasets (Al-Khalifa et al., 2026a). This gap persists even in multilingual efforts, such as (Srinivasan and Choi, 2022), which constructed a large-scale politeness dataset across multiple languages but did not include Arabic. Consequently, prior Arabic NLP research has predominantly focused on related tasks such as sentiment analysis (Al Motairi and Hadwan, 2024), offensive language detection (Mubarak et al., 2021; Albalawi and Yafooz, 2025), and hate speech detection (Alhazmi et al., 2024).

One of the few studies addressing politeness classification in Arabic is (Al-Khalifa et al., 2024). The authors investigate the effectiveness of transfer learning and large language models (LLMs) for politeness classification in Arabic social media text. Using a manually annotated dataset of 500 tweets from X, they compare fine-tuned Arabic language models (MarBERT and CamelBERT) with LLMs (GPT-4o-mini, Cohere Command, and JAIS 30B Chat) under zero-shot and few-shot prompting settings. The findings indicate that LLMs, particularly JAIS, outperformed the fine-tuned models, achieving an F1-score of 70.87%. The study is further extended in (Al-Khalifa et al., 2026b), which introduces a larger dataset comprising 10,000 annotated samples and evaluates a broader range of models. Among these, MARBERT achieved the best overall performance, with an accuracy of 0.9119 and a macro-F1 score of 0.8582, outperforming all other evaluated models.

3. System Overview

3.1. Subtask A: Politeness Classification

We experimented with multiple transformer-based models, including multilingual models (XLM-R and XLM-R Large) and Arabic-specific models (CAMeLBERT, MARBERTv2, ARBERTv2, and GATE AraBERT v1). These models were fine-tuned individually on the provided training data to predict the politeness label.

To improve robustness and performance, we employed a weighted soft-voting ensemble that combines class-probability outputs of the individual models. Specifically, for an input text, each model produces a probability distribution over the politeness classes. Each distribution is multiplied by a predefined weight, and the weighted probabilities are summed across models. The final prediction is obtained by selecting the class with the highest combined probability.

The model weights were determined using a grid search on the validation set. Candidate weights ranging from 0.0 to 1.0 (with increments of 0.1) were evaluated under the constraint that their sum equals 1 (3,003 combinations). For each weight combination, we computed the macro-F1 score on the validation set, and the configuration achieving the highest macro-F1 score was selected for the final system.

The optimal weight configuration was: MARBERTv2 (0.20), CAMeLBERT (0.10), AraBERTv2 (0.20), XLM-R (0.10), and XLM-R Large (0.40). The GATE AraBERT v1 model was assigned a weight of 0.0 and was therefore excluded from the final ensemble.

3.2. Subtask B: Category Prediction

Our approach consists of three main stages: data preprocessing and normalisation, MARBERT-based multi-label classification, and threshold-based post-processing.

3.2.1. Data Cleaning and Normalisation

We applied a comprehensive preprocessing and normalisation pipeline to improve data consistency and reduce noise in Arabic social media text. First, we removed Arabic diacritics and elongation characters (tatweel), which do not contribute to semantic meaning but increase lexical variability. We then normalised orthographic variants of Arabic letters to a canonical form, including mapping different

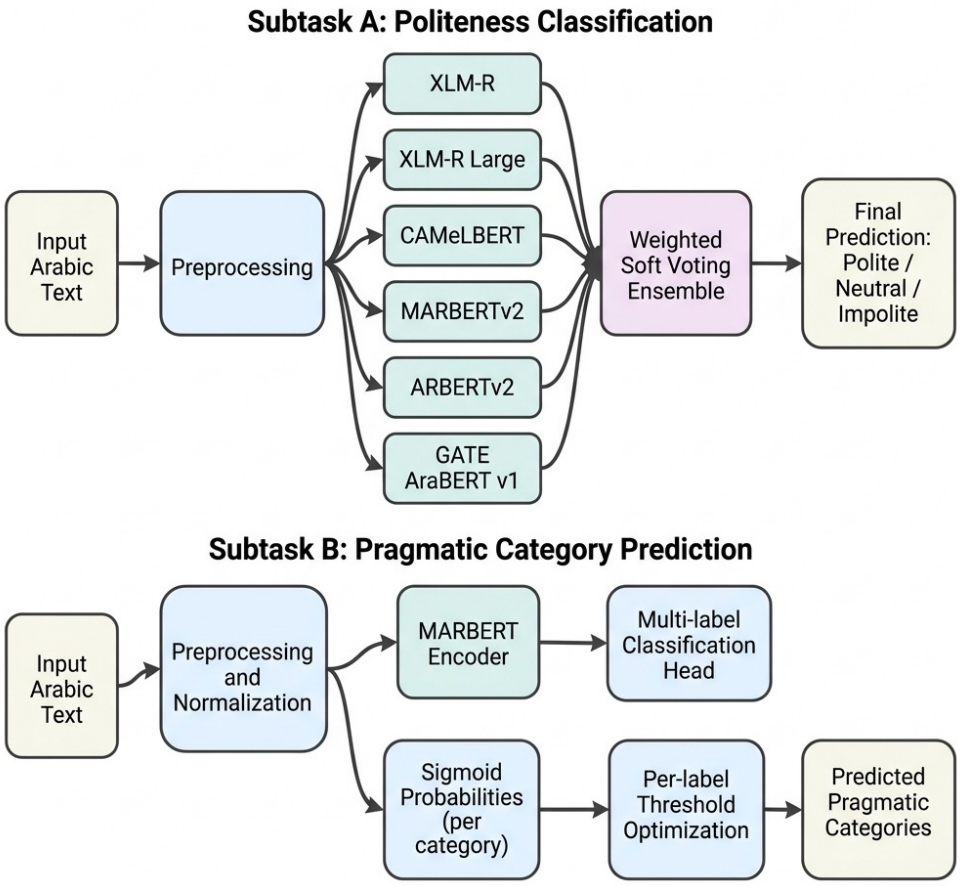

Figure 1: Overview of the proposed systems for the AdabEval 2026 shared task. The top pipeline shows the ensemble-based approach for politeness classification (Subtask A), while the bottom pipeline illustrates the MARBERT-based multi-label classification with threshold optimisation for pragmatic category prediction (Subtask B).

forms of Alef (إ، أ، آ) to a standard Alef (ا), converting Yeh variants (ى، ئ) to (ي), normalising Waw variants (ؤ) to (و), and converting Teh Marbuta (ة) to Heh (ه). These transformations reduce sparsity caused by orthographic variation.

To address informal writing and exaggerated expressions common in social media, we removed character elongation by collapsing repeated characters into a single instance (e.g.,تكفووون → تكفون). We also removed URLs and user mentions, which do not contribute to pragmatic classification and may introduce noise. Finally, whitespace was normalised by removing extra spaces and ensuring consistent token separation.

In addition to text normalisation, we standardised pragmatic category annotations using a predefined mapping to canonical label names. The dataset contained variations in category labels due to differences in formatting, bilingual annotation (Arabic and English), and inconsistent spacing. We mapped all variants of each category to a single normalised label (e.g., “Thanks & gratitude - الشكر و الامتنان” and “الشكر و الامتنان” were both mapped to *Thanks_and_gratitude*). When a direct mapping was unavailable, fallback normalisation rules were applied to extract the canonical English label and convert it into a consistent token format.

These preprocessing and normalisation steps improved label consistency, reduced vocabulary sparsity, and enhanced model performance by providing cleaner and more uniform input representations.

3.2.2. Model Architecture

Our system is based on MARBERT (Abdul-Mageed et al., 2021), a pretrained transformer model designed for Arabic social media text. MARBERT is built on the BERT architecture and trained on large-scale Arabic social media corpora, making it well suited for capturing informal language patterns and pragmatic cues. Given an input text x, MARBERT produces a contextualised representation, which is passed through a linear classification layer to gen-

erate logits for each category. Since subtask B is a multi-label classification task, we apply a sigmoid activation function independently to each category to produce probability scores.

3.2.3. Training and Threshold Optimisation

The MARBERT model (Abdul-Mageed et al., 2021) was fine-tuned on the provided training data using the HuggingFace Transformers library. During inference, the model outputs probability scores for each pragmatic category. To convert these probabilities into binary predictions, a decision threshold must be defined. Instead of using a fixed threshold (e.g., 0.5), we optimise label-specific thresholds on the validation set.

Specifically, for each category, we select the threshold that maximises its F1-score. This choice is motivated by the fact that threshold selection directly controls the trade-off between precision and recall, which significantly impacts classification performance (Lipton et al., 2014). Since the official evaluation metric is macro-F1 and the dataset is imbalanced across categories, optimising thresholds with respect to F1 is particularly appropriate (Sokolova and Lapalme, 2009).

Prior work has shown that tuning decision thresholds to maximise F1-score is effective in both binary and multi-label classification settings (Lipton et al., 2014). Therefore, we adopt per-label threshold optimisation to better capture category-specific characteristics.

3.2.4. Post-Processing and Prediction

At inference time, the model produces probability scores for each pragmatic category. We convert these scores into binary predictions using the optimised label-specific thresholds obtained from the validation set. A category is predicted if its probability exceeds the corresponding threshold.

This approach allows the model to account for differences in label distribution and classification difficulty across categories, particularly for low-frequency classes.

4. Experimental Setup

4.1. Data Splits

For subtask A, during the development phase, we augmented the training data by combining the training sets from subtask A and subtask B, as well as the validation set from subtask B. This strategy increased the amount of available training data and improved model generalisation. The validation set from subtask A was reserved exclusively for hyperparameter tuning, model selection, and ensemble weight optimisation.

For the final submission, we trained the models using the combined training and validation data from both subtask A and subtask B. The official subtask A test set, with hidden labels, was used for evaluation through the shared task leaderboard.

For subtask B, we followed the official dataset splits provided by the shared task organisers. The training set was used for model training, the validation set was used for hyperparameter tuning and threshold optimisation, and the test set was used for final evaluation.

4.2. Implementation Details

All experiments were implemented using the HuggingFace Transformers framework (Wolf et al., 2020) and PyTorch. We used pretrained transformer models that were fine-tuned on the provided dataset using GPU acceleration.

We used consistent training settings across all transformer models to ensure fair comparison and stable optimisation. The number of training epochs was determined separately for each model using early stopping based on validation performance.

All models were fine-tuned using the HuggingFace Transformers library with the AdamW optimiser and cosine learning rate scheduling. Mixed-precision training (FP16) was enabled when GPU resources were available to improve training efficiency. Tables 5 and 6 in the Appendix summarises the training hyperparameters used in our experiments.

4.3. Evaluation Metrics

The official evaluation metric used for leaderboard ranking is the macro-averaged F1-score. This metric assigns equal weight to each class or category, regardless of its frequency, making it particularly suitable for imbalanced datasets. Macro-averaged F1-score provides a balanced assessment of system performance across all politeness levels and pragmatic categories.

5. Results

Our systems achieved first place in both subtasks, demonstrating the effectiveness of transformer-based ensembling and MARBERT-based multi-label classification for Arabic politeness detection.

5.1. Subtask A: Politeness Classification

5.1.1. Validation Performance

Tables 2 and 3 present the validation performance of the individual transformer models and their ensemble. Among individual models, GATE AraBERT v1 achieved the highest macro-F1 score of 90.76%

and an accuracy of 85.59%. CAMeLBERT and ARBERTv2 also achieved strong performance, with macro-F1 scores exceeding 90%. The ensemble further improved overall performance by leveraging the complementary strengths of the individual models, resulting in improved robustness and generalisation compared to single-model approaches.

Model	Macro-F1	Accuracy
MARBERTv2	89.61	84.34
CAMeLBERT	90.91	85.37
XLM-R	88.74	80.21
XLM-R Large	88.46	81.84
ARBERTv2	90.04	84.92
GATE AraBERT v1	90.76	85.59
Ensemble	**91.63**	**87.19**

Table 2: Validation performance for subtask A.

Model	Neutral	Polite	Impolite
MARBERT	0.93	0.83	0.77
CAMeLBERT	**0.94**	**0.86**	0.76
XLM-R	0.93	0.82	0.65
XLM-R Large	0.92	0.84	0.70
ARBERT	0.93	0.84	0.78
GATE AraBERT v1	**0.94**	0.84	0.79
Ensemble	**0.94**	0.84	**0.83**

Table 3: Per-category validation F1-scores for subtask A.

5.2. Subtask B: Category Prediction

5.2.1. Validation Performance

We evaluated our MARBERT-based approach on the validation set to optimise classification thresholds. The optimal threshold was determined through a grid search to maximise macro-F1 score. We optimised decision thresholds separately for each category using the validation set. The optimal threshold for each category was selected to maximise its individual F1-score. This per-label threshold optimisation resulted in a macro-F1 score of 0.728 on the validation set.

Table 4 presents per-category F1-scores on the validation set. Performance was strongest on frequent categories such as gratitude, greetings, and prayers. The lower performance on categories such as hospitality was primarily due to limited training examples.

6. Conclusion

In this paper, we presented our systems for the AdabEval 2026 shared task on Arabic politeness detection and pragmatic category prediction. Our approaches leveraged pretrained transformer models and ensemble learning to capture linguistic and pragmatic features in Arabic social media text.

Category	Per-label threshold	F1-score
Gratitude	0.10	0.898
Greetings	0.10	0.875
Prayers	0.10	0.839
Respect	0.10	0.794
Insult	0.30	0.789
Admiration / Love	0.10	0.716
Criticism	0.15	0.641
Hospitality	0.10	0.000

Table 4: Per-category validation F1-scores for subtask B.

For subtask A, we developed a weighted soft-voting ensemble combining multilingual and Arabic-specific transformer models. Our approach achieved a macro-F1 of 0.89 and 0.93 accuracy on the official test set, ranking first among all participants. The results demonstrate the effectiveness of ensemble learning in leveraging complementary strengths of diverse pretrained models.

For subtask B, we proposed a MARBERT-based multi-label classifier with validation-based threshold optimisation and post-processing. It achieved a macro-F1 of 0.58 on the official test set, also ranking first among all participants. The strong performance highlights the importance of domain-specific pretrained models and threshold optimisation for Arabic pragmatic category prediction.

Our results show that pretrained transformer models provide a robust framework for Arabic politeness detection, particularly when combined with ensemble techniques and threshold optimisation. However, performance on rare categories remains challenging due to limited training data.

In future work, we plan to explore data augmentation techniques, larger pretrained models, and instruction-tuned language models to further improve performance, particularly for low-resource pragmatic categories. We also aim to investigate more advanced multi-label learning approaches and prompt-based methods for pragmatic analysis.

7. Acknowledgments

We would like to thank the organisers of AdabEval 2026 for providing the datasets and organising this challenge.

8. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 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*, pages 7088–7105.

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.

Hend Al-Khalifa, Nadia Ghezaiel, Maria Bounnit, Hend Hamed Alhazmi, Noof Abdullah Alfear, Reem Fahad Alqifari, Ameera Masoud Almasoud, and Sharefah Ahmed Al-Ghamdi. 2026a. 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 (LREC 2026)*, Palma, Mallorca, Spain.

Hend Al-Khalifa, Nadia Ghezaiel, Maria Bounnit, Hend Hamed Alhazmi, Noof Abdullah Alfear, Reem Fahad Alqifari, Ameera Masoud Almasoud, and Sharefah Ahmed Al-Ghamdi. 2026b. Adab: Arabic dataset for automated politeness benchmarking–a large-scale resource for computational sociopragmatics. *arXiv preprint arXiv:2602.13870*.

Reem K Al Motairi and Mohammed Hadwan. 2024. Sentiment analysis methods for arabic content on social media: A systematic review. *Ingénierie Des systèmes D'information*, 29(1).

Amal Albalawi and WM Yafooz. 2025. Arabic offensive text classification using emojis: including emoji data in arabic natural language processing. *International Journal of Electrical and Computer Engineering*, 15(3):3332–3345.

Ali Alhazmi, Rohana Mahmud, Norisma Idris, Mohamed Elhag Mohamed Abo, and Christopher Eke. 2024. A systematic literature review of hate speech identification on arabic twitter data: research challenges and future directions. *PeerJ Computer Science*, 10:e1966.

Reem Alqifari, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Maria Bounnit, Hend Al-Hazmi, 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)*, Palma, Mallorca, Spain. Co-located with the 2026 International Conference on Language Resources and Evaluation (LREC 2026).

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.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on twitter: Analysis and experiments. In *Proceedings of the sixth arabic natural language processing workshop*, pages 126–135.

Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2024. Computational politeness in natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–42.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical*

methods in natural language processing: system demonstrations, pages 38–45.

9. Appendix

9.1. Hyperparameter Configuration Details

Parameter	value
Max sequence length	256
Tokeniser	AutoTokeniser from pre-trained model
Data collator	DataCollatorWithPadding
Batch size (train)	8
Batch size (eval)	16
Gradient accumulation	2
Effective batch size	16
Learning rate	3×10^{-5}
Epochs	Up to 15 (early stopping)
Optimiser	AdamW
Weight decay	0.01
Warmup ratio	0.06
LR scheduler	Cosine
Mixed precision	FP16 if CUDA available
Threshold optimisation	Argmax (no threshold tuning)

Table 5: Configuration for subtask A.

Parameter	Value
Max sequence length	192
Tokeniser	AutoTokeniser from pre-trained model
Data collator	DataCollatorWithPadding
Batch size (train)	10
Batch size (eval)	10
Gradient accumulation	1
Effective batch size	10
Learning rate	2×10^{-5}
Epochs	Up to 50 (early stopping)
Optimiser	AdamW
Weight decay	0.01
Warmup ratio	0.06
LR scheduler	Linear (default HuggingFace)
Mixed precision	FP16 if CUDA available
Threshold optimisation	Validation-based grid search

Table 6: Configuration for subtask B.